THE UNIVERSITY
of EDINBURGH

PROFILING AND DISAGGREGATION OF ELECTRICITY DEMANDS MEASURED IN MV DISTRIBUTION NETWORKS

Andreas Paisios

Doctor of Philosophy

The University of Edinburgh

2017

Abstract

Despite the extensive deployment of smart-meters (SMs) at the low-voltage (LV) level, which are either fully operational or will be in the near future, distribution network operators (DNOs) are still relying on a limited number of permanently installed monitoring devices at primary and secondary medium-voltage (MV) substations, for purposes of network operation and control, as well as to inform and facilitate trading interactions between generators, distributors and suppliers.

Accordingly, improved and sufficiently developed models for the analysis of aggregate demands at the MV-level are required for the correct assessment of load variability, composition and time-dependent evolution, necessary for: addressing issues of robustness, security and reliability; accomplishing higher penetration levels from renewable/distributed generation; implementing demand-side-management (DSM) schemes and incorporating new technologies; decreasing environmental and economic costs and aiding towards the realisation of automated and proactive "smart-grid" networks. The analysis of MV-demand measurements provides an independent source of information that can capture network characteristics that do not manifest in the data collected at the LV-level, or when such data is restricted or altogether unavailable. This information describes the supply/demand interactions at the mid-level between high-voltage (HV) transmission and LV end-user consumption and opens possibilities for validation of existing bottom-up aggregation approaches, while addressing issues of reliance on survey-based data for technical and economic power system studies.

This thesis presents improved and novel methodologies for the analysis of aggregate demands, measured at MV-substations, aimed at more accurate and detailed load profiling, temporal decomposition and identification of the drivers of demand variability, classification of grid-supply-points (GSPs) according to consumption patterns, disaggregation with respect to customer-classes and load-types and load forecasting. The developed models are based on a number of traditional and modern analytical and statistical techniques, including: data mining, correlational and regression analysis, Fourier analysis, clustering and pattern recognition, etc. The approaches are demonstrated on demand datasets from UK and European based DNOs, thus providing specific information for the demand characteristics, the dependencies to external parameters and to socio-behavioural factors and the most likely load composition at the corresponding geographical locations, while the approaches are also intendent to be easily adaptable for studies at equivalent voltage and demand aggregation levels.

Acknowledgements

This work would not have been possible without the support of my supervisor Dr Sasa Djokic. His help and guidance throughout the course of my studies have been essential and for that I am ever so grateful. I would also like to extend my gratitude to Dr Adam Collin for his help and advice, particularly during the earlier stages of my studies, to Dr Ioannis Papastathopoulos from the School of Mathematics, for his useful inputs on some of the methodological aspects of this thesis and to Mr Alastair Ferguson from Scottish Power Energy Networks, for the provision of data and his collaboration on published work. Thanks are also extended to Dr Daniel Friedrich and Prof Matti Lehtonen, who acted as my thesis examiners and who provided valuable and insightful feedback.

Special thanks to my mother, father and brother, who have always been there for me and who are always in my mind, despite the distances that separate us. Special thanks to Maria, who has shared this experience with me and who has always been beautiful, kind and understanding.

To Camilla, Damilola, James and Wei, your advice and encouragement, particularly during the last few stressful months is appreciated beyond words. I wish you all the best. To all my colleagues, staff and students at the Institute for Energy Systems and in King's Buildings, thank you for providing such a friendly working environment.

To all the great people that I have met during my time in Edinburgh, you have enriched my life and I will always remember this period with the warmest of emotions. To all my friends back home, our shared memories of sunny summer afternoons have kept me going.

This research was funded by the Engineering and Physical Sciences Research Council (EPSRC).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Andreas Paisios

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AC	Air Conditioning
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BSC	Balancing & Settlement Code
CDD	Colling Degree Days
CDF	Cumulative Distribution Function
CFL	Compact Fluorescent Lamp
DFT	Discrete Fourier Transform
DG	Distributed Generation
DNO	Distribution Network Operator
DSM	Demand Side Management
DTM	Dynamically Teleswitched Meter
E10	Economy-10
E7	Economy-7
EU	European Union
FA	Factor Analysis
FFT	Fast Fourier Transform
GB	Great Britain
GSP	Grid Supply Point
HDD	Heating Degree Days
HV	High Voltage
HVAC	Heating, Ventilation & Air Conditioning
I&C	Industrial & Commercial
ICT	Information & Communication Technology
IFFT	Inverse Fast Fourier Transform
IGZ	Intermediate Geography Zone
K-W	Kruskal-Wallis Test
LAR	Least Absolute Residual

LED	Light Emitting Diode
LI	Long Interruptions
LOESS	Local Regression
LSOA	Lower-Layer Super Output Area
LV	Low Voltage
MA	Moving Average
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MPAN	Meter Point Administration Number
MSOA	Medium-Layer Super Output Area
MV	Medium Voltage
NIALM	Non-Intrusive Appliance Load Monitoring
NILM	Non-Intrusive Load Monitoring
Ofgem	Office of Gas and Electricity Markets
OLS	Ordinary Least Squares
OR	Ordinary Residential
PCA	Principal Components Analysis
RLOESS	Robust Local Regression
RTP	Real Time Pricing
SI	Short Interruptions
SM	Smart Meter
SPEN	Scottish Power Energy Networks
SSE	Sum of Square Errors
SSG	Sum of Square Group Errors
SSL	Shiftable Static Loads
SST	Total Sum of Squares
ToU	Time of Use
TR	Total Residential
UK	United Kingdom

Nomenclature

A	Solar Azimuth Angle
AP	Atmospheric Pressure
C^o	Degrees Celsius
d	Day of the Week
E	Solar Elevation Angle
m	Meter (distance)
mb	Millibars (pressure)
^{o}	Degrees (angle)
P	Active Power
P_B	Base Active Power
P_D	Active Power Difference
P_E	Expected Active Power
pf	Power Factor
PF_R	Reference Power Factor
P_{NTH}	Non-Thermal Demand
P_{SR}	Active Power Seasonal Range
P_{TH}	Thermal Demand
Q	Reactive Power
R^2	Coefficient of Determination
r_{co}	Cut-Off Correlation Level
RH	Relative Humidity
r_{xy}	Pearson's Coefficient
$r_{xy}{}'$	Spearman's Coefficient
s	Second (time)
SI	Solar Irradiance
T	Temperature
t	Time (half-hour timestamp)
$T_{thres.}$	Temperature Threshold
V	Voltage
W	Watt

Chapter 1: Introduction

The environmental and socio-economic challenges imposed by climate change and fossil-fuel depletion are driving changes in various systems of human function and organisation. Inevitably, in the centre of this evolution lies energy and more specifically, in the context of this study, electrical energy.

Modern electric power systems are characterised by increasing levels of complexity, regarding both the actual setup of the electricity grids and their components, as well as the multi-parametric interactions between networks, consumers, economic agents, technological innovations and, in general, the physical and social environments in which these systems function. One of the central attributes of complex systems is the emergence of new properties and relationships that are not always accounted for or anticipated by prior knowledge and existing theory about their individual components [1].

The deployment and proliferation of automated control, monitoring and communication infrastructure, supply and demand side interventions, the significant increase in the use of renewable energy sources and the decentralisation of power production through distributed generation, are some of the modern developments from which increasing complexity arises. Therefore, as mentioned in [2]:

"... it will be important to design simulation systems that can accurately represent both the grid and the reaction of consumers, in order to predict the emergent properties of the system under a range of different conditions..."

In such circumstances, an inductive approach to research, based on the wealth of data generated at various domains of the electricity grid and relying on the actual-measured system responses, can be used in order to inform the formulation of adequate models.

This thesis presents methodologies for the analysis of electricity demands, as measured at the medium-voltage distribution level, and formulated for the specific purposes of: 1) profiling and decomposition with respect to all relevant temporal-scales exhibiting significant demand variability, 2) determining the extent of demand dependencies to exogenous variables and people's daily schedules at the corresponding scales, 3) disaggregation of demands according to customer-sectors and load-types and 4) forecasting medium and short-term demand variability. Despite the fact that the methodologies are demonstrated on particular datasets, these are intendent to be easily adaptable for studies at equivalent aggregation levels and can be considered computationally efficient and parsimonious.

1.1 Background

An electrical power system is most commonly expressed as the sum of four distinctive components: generation, transmission, distribution and end-user consumption. The first three components represent the supply-side infrastructure, which is developed for the purpose of meeting end-user demand. The general distinction between transmission and distribution networks regards the complexity of the systems, in terms of the number of connected nodes and their corresponding voltage levels. The prescribed voltages vary according to the network setup and operating conditions with no internationally agreed standards.

In Great Britain (GB), the transmission network consists of the high-voltage (HV) interconnections at 400 kV, 275 kV and 132 kV[1] [3], enabling movement of bulk electrical energy from the generation sides to the main demand centres. The HV transmission system is interfaced to the medium-voltage (MV) distribution networks through transformer substations at: 400:132 kV and 275:132 kV in England and Wales and 275:33 kV and 132:33 kV in Scotland [3], [4]. Then, primary distribution substations at 132:33 kV, 132:11 kV and 33:11 kV [5] and secondary distribution substations at 11:0.4 kV [6], provide the necessary voltage reduction and system expansion for delivery to end-users[2].

The traditional setup of electricity grids positions the distribution networks and network operators (DNOs in GB) in the mid-point between generation and consumption and the MV-substations as the physical hubs facilitating the high-to-medium and medium-to-low voltage level transformation and the subsequent distribution of electricity to customers. It is therefore not surprising that despite the control of generation and bulk supply at the HV-level and the increased proliferation of automated control systems at the LV-level, much of the system's operation, resource allocation and stability is a responsibility of agents at the distribution level. This responsibility extends to issues of continuity and reliability[3], while further pressures are added to the distribution networks from increasing demands for decentralisation, due to the fact that the traditional hierarchical topology of electrical grids is now considered to be outdated [7] [8]. Thermal efficiency of centralised power plants is between one third to one half [9]; losses from transmission accounts for 2 % (in GB) [10] and 8 % (world average) [11]; generation capacity generally exceeds real-time demand in order to accommodate the, less

[1] The 132 kV level is considered as distribution in England and Wales and as transmission is Scotland.
[2] A small number of large customers, e.g. heavy industry, are connected to the HV and MV networks
[3] In GB, reliability is shown to be exceedingly high at the HV-level, with the vast majority of faults and interruptions being located at the distribution MV and LV levels [4], [7], [8].

frequently occurring, peak demands [12]; while supply interruptions originating at the HV-levels result in domino-effect failures (although these are infrequent in GB) [13].

In addition, the challenges arising from climate change and fossil-fuel depletion mean that decentralisation through distributed generation (DG) is typically associated with renewable energy sources. This coupling creates further challenges, including: the intermittency of renewable energy sources and technical considerations of power quality (e.g. voltage raise, transients, harmonic distortions and frequency instability); economic viability (e.g. power price competition) and security of operation, as outlined in [14], [15] and [16]. This is particularly the case since the distribution networks were initially developed as passive systems with no anticipated bilateral flows [17] and therefore considerable effort is put towards grid modernisation. A detailed report on the progress to date and future plans to meet the 2050 goal of: " reduction of carbon dioxide emissions by at least 80 % from their 1990 baseline", as presented to the House of Commons, in the UK, can be found in [18]. The corresponding grid modernisation multi-year plan for the US can be found in [19]. It is therefore recognised that adaptation to the new operating conditions is not only an option, but rather a desirable path towards energy security and sustainability.

To this end, much of academic, industrial, legislative, market and social interest is focused on what has come to be known as the "smart-grid". This refers to the next-generation electricity grid, which can be broadly defined as: an automated power network, with decentralised modes of generation and control, enabled by bidirectional information and communication technologies (ICTs), facilitating real-time demand and supply balancing, while maintaining high levels of resilience and protection against faults, disturbances and cyber-attacks [13], [20], [21]. This ambitious project requires changes to the current systems, not only in generation and grid architecture, but also for enabling direct/indirect and dynamic consumer participation. Such measures are collectively referred to as demand-side-management (DSM) and although DSM initiatives have been proposed and implemented since the 1970s and 1980s [22], they have been restricted and discouraged due to technological limitations. The recent advances in ICTs have positioned the concepts of smart-grid and DSM in the same realm and the two are usually discussed together.

The central proposition of DSM states that it is more energy efficient and economically viable to influence suitable changes in electricity demand, than to constantly adapt supply levels to meet customers' requirements [23]. This can be accomplished, in the long-term, by introducing and popularising energy-efficient consumption habits and the corresponding use of energy-efficient appliances (e.g. LED-lighting [24]) and standards have been developed and

adopted in many regions, with perhaps the most familiar example being the EU "standards and energy labels for appliances" [25]. Concerns regarding systemic responses within the economy, such as the rebound effect, have also been raised, in some cases questioning the actual long-term benefits of some of the proposed measures, as discussed in [26], [27] and [28].

Apart from the long-term schemes, within a modernised smart-grid, DSM usually refers to short-term and preferably real-time demand modifications. Traditionally, these approaches have been restricted to large deferrable industrial loads which can, if necessary, be disconnected from the grid to restore supply-demand balance, as opposed to increasing supply using spinning generation reserves [29]. The requirements for such reserves can be reduced by properly shifting loads to off-peak periods thus increasing utilisation of generation [30] and, for the desired scale of impact, these flexibilities need to extent to the residential and commercial sectors [29]. DSM interventions are therefore aimed at modifying demand to produce load profiles that maintain a stable supply-demand balance [31] and will, ideally, be able to minimize electrical disturbances (e.g. maintain proper system frequency [32] such as +/- 1 % of nominal for the UK [33]).

The short-term and real-time DSM schemes can be broadly categorized as direct and indirect as well as market and non-market based, although there are extensive intersections between these categories, particularly when the proposed applications can be implemented only within a smart-grid enabled network. For example, static time-of-use (ToU) tariffs for encouraging load shifting of specific load-types (e.g. economy-7 night-time tariffs for electric storage heating in the UK [34]), can be considered as indirect (and non-intrusive). With the installation of smart-meters (SMs) however, these approaches are extended to flexible or real-time pricing (RTP), that enables tailored load-shifting according to customer and network needs [35]. Since load-shifting can be accomplished via price and tariff variations, these approaches involve market participation [36], while supply-demand balancing mechanisms are already determined through the interactions of generators, distributors, suppliers and consumers (e.g. UK settlement periods, outlined in [37]). Considerable effort is put for the operational optimisation of the interactions between all of these agents; based on constraint, heuristic, game-theory and otherwise formulated approaches, as in [23], [38], [39] and [40].

These automated and responsive power networks [41], based on the use of smart-metering and real-time DSM, have the potential benefits of: increased reliability and faster recovery after faults [2], improved integration of intermittent renewable energy sources [42] [43] [44], more appropriate conditions for real-time market transactions [45], incentives for network

investments, reduced costs for individual consumers and the economy, as well as the environmental and social benefits of decarbonisation. SMs are also considered a vital element for the widespread adoption of electric-vehicles, due to the limitations of the current distribution infrastructure. This is due to the anticipated increase in requirements for charging, which can be accommodated by an automated system that enables real-time electricity allocation, thus decreasing total stresses on the system [46] [47].

There is however an ongoing discussion about the realistic benefits, risks and challenges regarding a number of proposed or partly-implemented developments. For example, it has been argued that the massive SM deployment (e.g. plan for 50 million gas and electricity meters and full domestic SM coverage by 2020 in GB [48] and an estimated €51 billion cost for full SM installation in the EU [49]) is not justified by the projected environmental, economic and operational benefits and that the anticipated changes in customer consumption patterns are better enabled only with sufficiently developed time-of-day and real-time pricing [50]. The UK government, in the official online SM guide [51], describe the benefits of SM installation in terms of: 1) real-time information on consumption habits and therefore improved management of energy use by the consumers, 2) billing based on actual-consumption in contrast with traditional estimated consumption and 3) easier switching between suppliers. Therefore, apart from the increased flexibility of customer-supplier communication, the benefits presented are limited to customers altering their behaviour based on SM information about their consumption patterns.

This seems to present a gap between the "vision" of smart-grid networks, as discussed by academia and industry, i.e. in the context of technological modernisation of the grid that allows DG, renewable integration and a real-time proactive network, and the one discussed by governmental institutions. Assuming that networks and markets adjust to facilitate DSM interventions only, the benefits will be a result of "social-engineering", i.e. changes in customer behaviour responding to price/tariff variations and not a result of technical-engineering that can accommodate the projected increases in energy requirements[4] [52]. Furthermore, cost-benefit analysis suggests that improvements in network operation and power quality control requires a critical mass of SMs to be installed [53].

[4] Although decreasing electricity demand trends can be shown in some cases, e.g. recent decreasing trends for the UK [54]. Analysis of economic development and electricity consumption produces varying results, from conservation and neutrality to positive correlation and support of the growth hypothesis [55].

Most of the concern and criticism is, however, concentrated on issues related to data privacy and security. High-frequency metering offers a high-resolution view of people's schedules and habits and can be used to determine occupancy and energy use patterns for individual households. This increases the vulnerability against illegal activities in cases where the information is leaked or hacked (i.e. criminal activity by intersection of communications), discussed in [56] and [57]. Consumer groups have also raised concerns about the uses of SM data by commercial entities, for purposes of advertisement, insurance adjustment and targeted marketing [58] . Similarly, SM-data availability can potentially be exploited by governments and law-enforcement agencies, or by other parties for legal purposes. Cyber-security is of main concern since the ICTs can be borne to attacks for purposes of electricity theft and meter-manipulation, or malicious attacks on the electricity grid infrastructure by foreign or domestic adversaries [59] [60]. Furthermore, the growing body of research dedicated to the software-engineering side of LV-metering (e.g. optimization algorithms, communication and real-time response algorithms, development of protocols and standards, etc. [35], [38], [61], [62], [63] and [64]) implies a large-scale increase in complexity [65] [66] and thus an important challenge from a technical point of view.

It should be noted however, that despite the extensive research regarding, or aimed to address, the challenges of SM technologies and their integration to the economy, there appears to be limited academic-literature that can be considered as a direct criticism of the altogether deployment of these technologies. There appears to be a general consensus among the scientific community regarding the validity of arguments for the installation of SM devices and a recognition of the long-term opportunities it offers for grid-modernisation.

This thesis does not take a stance for or against any of the proposed developments but it is rather based on the assumption that grid modernisation is necessary even if only to meet increasing energy requirements and that it is inevitable, with or without ICTs at every level of the grid. The methodologies presented are therefore aimed at the retrieval of valuable information from already existing metering devices, installed at the distribution level of the grid. Feedback data from MV-substations corresponds to a degree of system aggregation that can be considered informative of the supply-demand interactions, yet not too complex as to be ambiguous, or too specific to individual consumers. It is descriptive of consumption characteristics at certain locations, sectors and temporal periods, less bound by privacy concerns and more accessible by statistical analysis due to the aggregation of data; and positioned at the centre of some of the ongoing and anticipated changes in power networks.

1.2 Thesis Scope and Structure

This thesis concentrates on methods for the analysis of measurements (i.e. active power - *P*, reactive power - *Q* and voltage - *V*) from, generally, already available metering apparatus at the MV-level and investigates the possibilities for retrieving useful information that can facilitate improved system operation and control, implementation of DSM interventions and inform planning and development considerations. Each chapter is dedicated to particular aspects of the analysis and the thesis is organised so that more basic results (e.g. load-profiles) precede more complex and involved methodologies (e.g. load-disaggregation). However, the results and conclusions of earlier chapters are often used for the analysis in subsequent chapters. The thesis structure can be summarised as follows:

In **Chapter 2**, a review of existing literature is presented, putting the current study into context. This is structured in sections discussing existing literature relative to specific aspects of the thesis, corresponding to the analysis and results presented in Chapters 3 to 7. A brief overview of the statistical approaches used is also presented, although more detailed discussions are included in the following chapters, when these methods are explicitly used.

In **Chapter 3**, the demand datasets used in this thesis are presented. The variabilities of parameters *P*, *Q* and *V* are investigated through Fourier decomposition. Detailed profiling is presented with respect to the daily, weekly and yearly cycles and analysis of variance is used to determine periods of similar and dissimilar demand levels. The probability of occurrence of maximum (peak) and minimum demand levels and the benefits from profiling the rate-of-change of demands are also investigated.

In **Chapter 4**, the main meteorological drivers of short and medium term demand variability are determined through correlation and regression analysis, conducted for different temporal-scales. The analemma variables (solar azimuth and elevation angles) are introduced and evaluated as explanatory variables. For the same temporal scales, the statistical relationships between the *P*, *Q* and *V* parameters are also investigated. A detailed discussion is provided based on the analysis of residuals for the purposes of explaining certain features of seasonal demand variability, as well as for justifying the use of multiple-regression in the following chapters and the sensitivity of demand to weather conditions, throughout the year, is evaluated, based on a moving-window regression approach.

In **Chapter 5**, MV-substations are clustered according to average demand patterns. Then, based on a number of selected metrics and data retrieved from official consumption statistics, a methodology is presented for estimating percentages of contributions from different

customer-classes. The consistency of the resulting percentages as estimated from different metric-combinations is evaluated and the results are compared to the clustering classification. Initial results based on a model developed for generating diurnal demand profiles according to selected contributions from the domestic sector are also presented.

In **Chapter 6**, loads are initially decomposed into base, intermediate and peak demands, defined in terms of: annual base, daily base, daily variable but seasonally constant, and seasonally variable portions. Approaches are presented for determining temperature and solar irradiance threshold values, marking the periods when electrical heating, cooling and lighting loads are switched on. Two disaggregation approaches are investigated, relying on multiple-regression analysis and data-transformations according to the *P-Q* relationship (i.e. power factor), and results are presented for the diurnal and seasonal contributions of generic and specific load-types to the total measured demand.

In **Chapter 7**, a dynamic multiple-regression approach for forecasting short-term and medium-term active and reactive power demands is presented. The method does not rely on specifying predictors prior to the analysis, but rather on evaluating the modelling performance of various combinations of predictors before applying the selected pairs for forecasting. The use of the analemma variables for active and reactive power demand forecasting is also explored, as an alternative (or complementary) to the use of meteorological data.

Finally, **Chapter 8** presents a summary of the main findings and discusses the challenges, limitations, as well as proposed directions for further research.

1.3 Thesis Contributions

The work presented in this PhD thesis is indented to aid and inform further studies, either through the reproduction of the developed methodologies, or by direct consultation of the results, as presented. An effort has been made to combine demand datasets which correspond to shared or similar consumption characteristics and, in particular, results and approaches based on *P-Q* and *P-Q-V* records are specific to UK distribution networks (i.e. north of England and south/central Scotland). The main contributions can be summarised as:

- An investigation of the statistical characteristics and temporal variabilities of active power, reactive power and voltage, through detailed decomposition and profiling. As most of the existing literature is concentrated on either total-system or individual/aggregated LV-profiles (or composite and nationwide survey-based profiles) the current study presents demand profiles as derived from a large data-sample of MV-substations. The results are

therefore representative of actual aggregate consumption characteristics, specific to locations and DNOs and with realistic and diverse customer-class mixtures.

- An investigation of the medium-term and short-term dependencies of electricity demands to external parameters, at the diurnal and seasonal time-scales. These results can inform the variable selection process for subsequent studies, particularly when these are specific to locations of similar climate characteristics. The analysis also demonstrates the temporal-scales at which causal relationships between demands and external parameters can be established, as well as the different possibilities that arise from analysis at each of these temporal-scales, with respect to generality or granularity.
- The introduction and an investigation of the applicability of the solar analemma variables (solar azimuth and solar elevation angles), for purposes of demand profiling and load disaggregation and forecasting. The use of these parameters shows improved demand modelling performances, while being directly related to seasonal/weather changes and can be used for the described purposes in the absence of meteorological data. (Publication from this research: [67]).
- The development of a methodology for the classification of MV-GSPs and the subsequent disaggregation of total measured demands into percentage-contributions from various customer-classes. The results are based on consumption patterns accessible directly from the MV-measurements and can be used for the assessment of proposed DSM interventions, or to inform system planning, expansion and DG-integration considerations, as different customer-classes exhibit particular demand characteristics.
- The development of methodologies for the disaggregation of total measured aggregate demands into generic as well as specific load-type categories. This work also presents results for balance-point temperature and solar irradiance levels and estimations of base, intermediate and peak demands, according to constant and variable portions of total consumption. The results can inform DSM interventions, through determining deferrable portions of the load, or for studies of the efficiency of thermal load consumption through the disaggregation of total demands into heating/cooling components, as well as for balancing variable demand and generation from renewable/DG sources. (Publications from this research: [68], [69][5]).

[5] [68] was selected from conference papers and invited to be published as a book-section. This resulted in the revised version in [69], which includes additional results. Both publications include demand-profiling analysis, but the primary focus, in both cases, is on load decomposition/disaggregation.

- The development of a dynamic multiple-regression model for short-term and medium-term active power and reactive power demand forecasting, based on meteorological and analemma parameters, flexible and adaptable according to the data availability of predictor variables. (Publication from this research: [67]).

1.4 Terminology

For the purposes of this study and throughout this text, several terms are used interchangeably. In particular:

- Load, demand and consumption are used to refer to the measured or estimated parameters: active power – *P* and reactive power – *Q*. Load is also used to denote power consumption from specific devices or types of equipment (e.g. thermal-resistive loads, in load disaggregation) and in the same context, this load corresponds to portions of the total measured demand.
- Diurnal and daily refer to the period of 24-hours of one day, which, based on the resolution of the available datasets, is represented by 48 half-hourly timestamps. The terms seasonal, yearly and annual correspond to one calendar year (365 days), from 1st of January to 31st of December, unless stated otherwise, e.g. when only weekdays are used, or when holiday periods are excluded. Weekdays and weekends have the conventional meaning; Mondays to Fridays for the former, Saturdays and Sundays for the latter. Additional days from leap-years (i.e. 29th of February) have been deleted for consistency between subsequent years.
- Grid-supply-points (GSPs), buses and substations are also used interchangeably, and correspond to the points in the distribution networks at which data are recorded. The nominal voltage levels are not the same for all datasets and the appropriate clarifications are made in Chapter 3.
- Customer-sectors, customer-classes and customer-categories correspond to the portions of the total aggregate demands that can be attributed (known or estimated) to groups of end-users with similar activities and consumption patterns, e.g. from the residential, commercial and industrial sectors. Measured demands at the MV-level are usually constituted from different combinations of percentage-contributions from these classes and the total is therefore a customer-mixture. In the same context, domestic corresponds to residential and non-domestic to commercial and industrial. In Chapter 5, further clarifications are made, with respect to the presented methodology and data availability.

System Used: All results presented in this thesis are based on computations conducted on an Intel i7-2630 QM CPU at 2 GHz, operating with 64-bit Windows-7 and implemented in Matlab R2016a (academic use – offline) [70].

Chapter 2: Literature Review

The literature review is divided into six sections. The first five correspond to the work presented through Chapters 3 to 7, while the sixth section discusses the literature relative to the statistical approaches used in this thesis.

2.1 Load Modelling and Profiling

Power system models are used in a number of studies related to all parts and voltage levels of the grid, from generation, through transmission and distribution, to end-user consumption. Load models at the MV and LV distribution networks are important for planning, operation and market analysis, as well as for addressing issues of stability and control and for predicting future grid characteristics, particularly important due to the introduction of new technologies in appliances and the integration of DG. Load models are also important for the success of DSM schemes, since they provide crucial information about consumption patterns and the availability of deferrable loads. A number of different definitions are found in literature for the description of load models and their uses.

First distinctions are usually made between static and dynamic load models. These correspond to the mathematical representations of electrical load parameters, appropriate to specific load-categories due to their shared characteristics (e.g. resistive loads, CFLs, LCD-displays, rectifiers, etc.). Static (or steady-state) load models describe active and reactive power demand dependencies on voltage and frequency changes, at specific time instances, as instantaneous step-changes with no transient characteristics. Dynamic load models are more appropriate for the analysis of dynamic (transient) system responses and load disturbances and do not assume instantaneous transfer from one state to the next [71]. Some common mathematical models used are the: constant power - *PQ*, constant impedance - *Z*, exponential, polynomial - *ZIP* and many others [72]. Based on the selected methods and used analytical/mathematical descriptions, two main approaches can be distinguished for providing the data-values necessary to build these models: the component based (bottom-up) and the measurement based (top-down).

Bottom-up based approaches are formulated from the combination and aggregation of individual load components, usually corresponding to specific appliances or load-categories. The primary source of information regarding load characteristics for these components is taken from experimental results, such as in [73], [74] and [75]. Individual components are frequently categorised into larger sets, a process which simplifies the methodologies and reduces

unnecessary complexity, in the sense that these categories correspond to similar-purpose appliances, e.g. "power-electronics" includes laptops and ICT loads, lighting includes incandescent lamps and CFLs, etc. [76]. The total load is a function of selected percentages for the individual load categories, according to the desired model specifications, i.e. aggregation for domestic consumption would correspond to different percentage-contributions from the same load categories as compared to the aggregation of load categories in commercial consumption. Aggregation aimed at representing bulk-supply-point loads includes specification of percentages according to load categories, as well as according to sector combinations, also known as customer-classes, typically: residential, industrial and commercial. This information is usually derived from survey-based socio-economic data, i.e. patterns for occupancy levels, time-varying probability of use for different appliances, sector-mixtures at different locations, etc. [77]. Examples of approaches for generating composite models can be found in [76], [78], [79], [80] and [81].

Top-down or measurement-based load modelling, relies on deriving load parameters from data recorded by different monitoring devices installed in the grid. These can measure the system-states under normal operating conditions, or they can be used for recording disturbances, typically regarding power quality and fault-events' detection [82], [83]. For power-flow quality and stability studies, transient states, regarding the changes in voltage/frequency and active/reactive power demands, are modelled as in [84], [85]. When the measuring point corresponds to higher levels of system aggregation, the parameters for individual loads are estimated based on various optimisation methods, such as in [86] and [87].

Measurements from individual customers and from various grid aggregation levels, can also be analysed statistically, e.g. to model the variability of demands, for purposes of determining the influence of external parameters and for demand forecasting, as well as for estimating availability of loads for DSM. This can be done for specific locations and can correspond to particular load sectors and/or load-categories. The resulting patterns are also used for electricity trading in the scheme of settlement periods. In this context, load models are more frequently described as load profiles (also referred to as consumption or demand profiles). These are the numerical and/or graphical representations of electricity demands, over selected time-scales and corresponding to aggregated or specified loads. There is no clearly defined distinction between load models and load profiles but, generally, load models consider the mathematical expressions describing the static or dynamic states of electrical parameters that can provide information on *P/Q* demands as a function of, typically, voltage and frequency changes, whereas load profiles aim to describe the variability of demands as a time-series

evolution, with less consideration given on the interactions between these parameters. The distinction is, therefore, between approaches of analytical and of statistical nature. The variability can be presented over different temporal-scales (i.e. daily, weekly, monthly, seasonally, annually and over-annually) and can correspond to different levels of load aggregation such as: LV-consumption from individual households, loads from MV-distribution feeders or as recorded from the HV-transmission networks.

In GB, Elexon [88] defines eight different load profiles [89], two of which are residential/domestic (unrestricted and economy-7) and six which are non-domestic (unrestricted, economy-7 and four more categorised according to peak-load factors, from <20 % to >40 %). The daily load profiles are constituted of 48 half-hourly settlement periods and these can be used to form settlement-days and settlement-years, corresponding to the average customer in each profile. The profiles are build form half-hourly measurements of individual customers and based on a sample selection process that takes into account population distributions per consumption range and within stratums according to GSP-groups in England, Wales and Scotland. The data analysis involves weighting per stratum and per GSP-group and averaging based on the population in each group. Regression analysis is used to determine coefficients for temperature and sunset-times and dummy variables are used for the days of the week. Evaluating the estimated relationships gives load profiles that correspond to the consumption of an average customer, in each profile-class, under the specified conditions and it is therefore possible to estimate loads in the absence of continuous half-hourly metering. A detailed description of the methodology is given in [89].

Electricity suppliers need to assess customer demand for electricity in advance for trading according to settlement periods, a process in which contracts can be made up to an hour before the actual dispatch. Imbalances between estimated and actual demands are handled by the system operator, i.e. National Grid in GB, in accordance with the balancing and settlement code (BSC). Imbalances have economic costs on suppliers/generators and therefore accurate forecasting is essential. A detailed description of the trading and balancing arrangements and the interactions between the various agents involved is given in [37]. In the above scheme and due to the balancing requirements, the necessity for producing accurate end-user profiles is central.

However, operation and control of the distribution networks, which is a responsibility of DNOs, requires load profiles at an aggregate level and preferably at the various levels at which DNOs function. Furthermore, while tariff arrangements for individual customers are determined through the eight Elexon profiles (in the absence of smart-metering), electricity

suppliers trade in bulk and with respect to the locations in which they operate. Distribution networks are therefore obliged to allow non-discriminatory access to all parties wishing to use information about their networks in order to facilitate their trades, including suppliers, as well as generators and within the regulatory frame set by the BSC. Sufficiently developed load profiles, derived from aggregated datasets, are therefore important for technical and operational decision making processes. In [90], authors compare the aggregate demands from a particular GSP with the estimated aggregated demands based on the number of customers (in the corresponding area) and according to the Elexon profiles; and find an imbalance between the two, discussing the possibility that the profiles might be considered outdated due to the introduction of DG and of new appliance-types. Therefore, detailed profiling of demands at the distribution level also opens possibilities for validation and improved formulation of existing customer profiles.

Chapter 3 presents a detailed load profiling approach based on data acquired from four different distribution networks and the respective geographical locations in which they operate. Fourier analysis is utilised to determine the various modes of temporal variability (i.e. time-series frequency decomposition) for active power, reactive power and voltage. Fourier transforms for load modelling and harmonic analysis are outside the scope of this thesis and the discussion is concentrated on the applications with respect to load profiling. Fourier analysis has been used before in this context, but primarily for purposes of demand forecasting as in [91] and [92]; and based on analysis of total-system demands. In [93], Fourier analysis is used to characterise electricity demands, but limited to the domestic sector, based on aggregate measurements from individual dwellings. In [94], Fourier decomposition is used to characterised daily, intra-daily and yearly demand cycles and it is subsequently combined with regression and autoregressive models, in order to generate synthetic profiles. The analysis is, however, concentrated on active power only and it is based on aggregate system demands. In [95] Fourier analysis is used for disaggregating HVAC (heating, ventilation and air-conditioning) consumption in individual commercial buildings.

In this thesis, no prior assumptions are made about expected periodicities and the inputs are raw datasets with no preliminary filtering applied. Furthermore, the data comes from a diverse set of GSPs corresponding to residential, commercial and industrial consumptions and discussions are provided regarding the different modes of temporal variability according to customer-mixtures. The success of reconstructing (synthesising) electrical parameters is also presented and shows distinctions according to the modelled parameters (i.e. P, Q and V). While it is generally assumed that different demand levels exist as a function of the hour of the day,

day of the week and day of the year, the analysis offers a detailed profiling of the periods of similar and dissimilar demand levels, determined through hypothesis-testing and based on analysis-of-variance models. Furthermore, and as discussed in [96], there are significant differences between the load profiles and the corresponding DSM availability between individual households and aggregate demands. The analysis in this study is based on a large sample of MV-substations and occurrence of peak/minimum demands is presented for the corresponding aggregation level. Variability in the occurrence of peak demands is shown with respect to the diurnal and seasonal components, which is important for the development of dynamic approaches to DSM. Such differences have been discussed before, but regarding specific customer-classes such as in [97], for the domestic sector and in [98], for the non-domestic sectors. In reality however, the majority of MV-substations supply diverse mixtures of these customer-classes.

2.2 Drivers of Demand Variability

The analysis of correlations and dependencies (Chapter 4) builds on the results of profiling (Chapter 3) in the sense that, while profiling determines the principal modes of variability, analysis of dependencies aims to determine the causes of this variability. This is one of the most extensively studied areas in power systems as the results are, primarily, utilised to build successful forecasting models, necessary for optimal operation and planning. In this context, a forecasting model based on a multiple-regression approach is presented in Chapter 7. In addition, the results of the dependency analysis are used for purposes of classification and disaggregation in Chapters 5 and 6.

In the long-term, economic development and fuel-prices, socio-political conditions, demographic trends and population distributions, technological advancements and climate change, have been demonstrated to have the most significant effects on electricity demand, as discussed in [99], [100], [101], [102], [103], [104] and [105]. This thesis is concentrated on drivers of medium-term and short-term demand variability, which can be defined within the range of one calendar year, from seasonal changes and down to the frequency of available data, i.e. half-hourly resolution.

Meteorological conditions have been shown to have the most significant effect on seasonal demand variability (i.e. medium-term), as discussed in [106], [107], [108] and [109]; particularly due to changes in temperature and solar irradiance levels. This is confirmed by the results of the current study. Other weather parameters such as relative humidity show only weak correlations and it is assumed that this is due to the climate characteristics corresponding

to the locations from which demand data is retrieved. However, authors in [108] demonstrate improved modelling performance when humidity is included, for UK data, and based on a monthly resolution. A detailed discussion on the effects of relative humidity is provided in Chapter 4. Other authors discuss the effects of wind speed, such as in [110], in terms of increased ventilation during warm periods, or in [111] and [112], in the context of increased wind power output. In this study, wind speed has been shown to have little to no effect on demand variability, although the effects of renewable-DG have not been investigated. The seasonally-constant portions of demand are discussed in Chapters 5 and 6, as these are shown to be primarily determined by customer-class and load-type compositions.

The non-linearity in the relationships between electricity demand and external variables (particularly temperature) have been identified and discussed by many authors, e.g. in [113], [114] and [115]. In [115], a comparison of the electricity demand-temperature relationship is presented among 15 European countries, indicating significant differences according to geographical location, a result which is attributed to higher demand for air-conditioning in southern European countries and similarly, higher demand for heating in northern Europe. Such distinctions in the demand-temperature relationship are further discussed in the context of disaggregation and for defining threshold temperatures in Chapter 6. The effects of multiple parameters are also recognised, as in [116], [117] and [118], including weather and socio-economic variables. In the context of multiple-regression, this thesis presents a detailed analysis of the particular aspects of the relationships between weather variables and electricity-demand, utilising data-filtering and analysis of residuals and discusses the possibility that some of these effects are psychological in nature (Chapter 4). The improved performance of models with more than one predictors is, nevertheless, widely recognised.

The use of the solar analemma parameters as explanatory variables in correlation and regression analysis with electricity demand is also explored (Chapter 4) and subsequently included in the forecasting models (Chapter 7). Closely related variables, such as the sunrise/sunset times have been used before, e.g. in regression analysis for determining the Elexon profiles in [89], but to the best of the author's knowledge, solar elevation and solar azimuth angles have not been used in the same way as in this thesis. Analemma variables are, however, frequently used in evaluating solar energy potentials, as in [119] and [120]. Further information is provided in Chapter 4.

The time-dependent sensitivities of electricity demand to external parameters are determined based on a moving-window (or rolling-window) regression (Chapter 4), an approach that can be considered as an alternative to determining regression coefficients per month of the year,

or per season, as it is the common practice. This seasonal segmentation is based on the demonstrated differences between demands during various periods of the year, both for the domestic and non-domestic sectors, and the results are reflected in profiling studies, e.g. in [97] and [98]. However, a continues parameter estimator provides information on the changes in the relationships throughout the year. A moving-window approach is used by authors in [121] and [122] for purposes of short-term electricity forecasting and based on a 2-dimensional window of length of up to 2-days and 2-hours before. A time-varying regression model for forecasting is also used in [123], demonstrating the efficiency of allowing regression coefficients to vary according to seasonal changes. The authors also discuss the importance of separating the models according to days of the week and hours of the day, an approach that is adopted in this thesis as well.

As discussed in Section 2.1 and within the context of load modelling, the relationships between load parameters have been extensively studied and usually expressed as analytical equations in static/dynamic load models. In contrast, this thesis presents a statistical analysis of load parameters (*P, Q* and *V*), based on diurnal and seasonal correlation/regression analysis. The resulting profiles for the time-dependent correlations between *P* and *Q* are subsequently used for load-disaggregation in Chapter 6.

2.3 GSP-Classification and Customer-Class Disaggregation

Classification of electricity load measurements according to end-user consumption patterns offers important information to distribution network operators and electricity suppliers. It facilitates demand-supply balancing and DSM schemes for specific geographical areas and according to particular sector-mixtures, as different customers have different seasonal and diurnal demand requirements and peak demands for these categories occur during different periods. Knowledge of the contributions of each category to the total demand can be used during the assessment of network DG-integration and for further planning considerations. Classification also enables more diverse tariff formulations and thus improved customer service. The technological developments in appliances and the installation of LV-DG can also affect load patterns and render existing profiles outdated. This area of study is therefore closely related to load profiling and, as in the case of load profiling, distinctions can be made between classification of individual customers and classification of aggregate measurements.

Classification of individual customers is being extensively studied, particularly due to the availability of LV-measurements through the installation of metering and smart-metering devices. The results are primarily intended to aid tailored and more diverse tariff formulations,

as well as to enable DSM interventions specific to each customer's consumption patterns. Examples of these approaches can be found in [124], [125], [126] and [127]. In this thesis (Chapter 5), the effort is concentrated on classification and customer-class disaggregation based on measurements retrieved at the MV-distribution level.

Detailed information about the load composition with respect to customer-sectors is generally not available to DNOs, apart from what can be found in the form of national and sub-national statistical data, as in [128]. This information however corresponds to regional and local-authority levels of aggregation, such as the lower-layer-super-output-areas (LSOAs) and the middle-layer-super-output-areas (MSOAs) in England and Wales, or as the intermediate-geography-zones (IGZs) in Scotland. Here, a problem arises due to the fact that there is no exclusive correspondence between the areas defined by the LSOAs, MSOAs or IGZs and the areas supplied by each MV-distribution feeder. In other words, a particular local authority area might be supplied by more than one MV-substation and similarly, a MV-substation might be positioned at the intersection of two or more local-authority areas.

In [111] a load-identification approach is implemented based on a similar set of MV-data as the one used in this thesis (i.e. Scottish substations). The author however relied on assumed typical load curves for residential, commercial and industrial consumption and classified a set of MV-substations according to a goodness-of-fit measure, estimated between unknown and typical load-curves. In [129] authors used clustering and Euclidean distance approaches to determine sector percentages for five unknown MV-substation profiles, based on a training dataset of 54 MV-substations, relying on similarity measures between per-unitised daily demand curves. In [130] authors cluster MV-customers according to normalised indices, which are however homogenous with respect to the customer-type and therefore no distinctions are made between different sectors. Similarly, in [131] authors assign load profiles to library-types using k-means clustering and inferences are made about the typical consumption patterns (i.e. residential and commercial), but these are not decomposed into specific percentages. In a bottom-up approach, authors in [132] determine representative load profiles for residential, commercial and industrial customers which are subsequently used to construct aggregate consumption patterns.

This thesis presents a classification of MV-substations based on average diurnal patterns, which is a common practice, however, the second stage of the analysis presents a customer-class disaggregation method. A number of different metrics are used to model known percentage mixtures (calculated from IGZ data) and the successful combinations of metrics are used to determine these percentages for the complete set of available MV-substations.

There is, therefore, no reliance on expected load-profiles as the relationships are directly determined from measurements. The diverse set of analysed metrics includes normalised diurnal profiles with respect to mean-demands, but also with respect to the range of seasonal variations and the final percentages can account for domestic and non-domestic customers, as well as for economy-7 residential and ordinary-residential customers.

Regarding the classification approach, a wealth of literature is dedicated to various clustering and pattern recognition techniques, within the realm of electrical power systems, but also for studies in artificial intelligence, economics, biology, social sciences, etc. It should be noted that, strictly speaking, classification and clustering are not equivalent, as the first presupposes existing classes determined by the research objectives, while the latter refers to an unsupervised partitioning of data. The more widely used clustering methods include connectivity models, such as: hierarchical clustering, nearest-neighbour, etc. and partitioning models, such as: k-means and follow-the-leader. Other methods employ distribution models, fuzzy logic clustering, data mining techniques, self-organising maps and artificial neural networks (ANN). In many cases, various techniques are combined to produce clusters based on the strengths of the individual approaches, facilitating validation and addressing computational efficiency. Reviews and comparisons of clustering and classification approaches for electricity customer-profile studies can be found in [133] and [134].

The clustering algorithm presented in Chapter 5 was developed specifically for this study and did not rely on any pre-existing methods. However, consultation of existing literature showed that it can be categorised as an agglomerative (bottom-up) hierarchical clustering approach [135], relying on correlation coefficients as metrics of similarity. Although no comparisons between this algorithm and existing hierarchical algorithms in terms of performance are presented in this thesis, a comparison between the final clusters and the percentages determined through the customer-class disaggregation procedure are presented.

2.4 Load Disaggregation

Load disaggregation generally refers to approaches developed for the separation of total measured electricity consumption into distinctive components, representing either specific appliances or load-categories with shared electrical and/or demand characteristics. A related but not equivalent concept is load decomposition and describes the process by which the time-series representation of total load can be separated into specific modes of temporal variability, as in [99], and [136] or as in the Fourier decomposition presented in Chapter 3, in order to give

some understanding of the demand structure and facilitate further studies such as load forecasting.

Decomposition approaches can, however, have applications for load disaggregation since the variability of total demand is usually associated with specific load-categories. For example, authors in [137] discuss the availability of deferrable loads for DSM and make distinctions between shiftable-static-loads (SSL) and seasonally variable thermal loads, i.e. for heating and cooling. Load categories suitable for direct load-shifting are often comprised of static and seasonal portions, meaning that estimation of the DSM potentials needs to take into account both components, e.g. cold-appliances are generally considered as static but have seasonal variability [106]. Similarly, decomposition approaches can be used for determining base, intermediate and peak portions of total demand, which do not correspond to specific load-categories, but are nevertheless useful for system operation and generation planning. This is usually approached through load-duration curves, as in [138], while authors in [139] present a statistical method based on cluster-analysis. In Chapter 6, a decomposition of total demand (from MV-substations) is presented in terms of the: seasonally-variable portion (which includes annual and daily peaks), hourly-variable portion (but seasonally-constant, which can be considered as the intermediate part) and two different base loads, i.e. the per-day base (seasonally-variable) and the global-base (minimum recorded demand).

Regarding load disaggregation, the vast majority of approaches are concentrated on its applicability to LV-measurements and particularly at the individual household or building level. This problem is considered to be an engineering and computational challenge, which is addressed in order to fully realise the benefits of SM deployment. While working models and advanced prototypes are available [140] and have been since the 1980s, it is an open research area due to the potential for improvements and the need for adaptability to the characteristics of new appliances. The necessity for LV-disaggregation arises from the fact that while SMs enable real-time consumption monitoring, the information regarding the consumption patterns of individual appliances is often unavailable, as it is included in the composite/metered signal. Disaggregation aimed at deconstructing this signal is referred to as non-intrusive-load-monitoring (NILM), or non-intrusive-appliance-load-monitoring (NIALM). The alternative is a direct monitoring of individual devices, which is considered costly and inconvenient and it is referred to as intrusive load monitoring. The privacy concerns discussed in the context of smart-meters in Section 1.1 are therefore even more relevant, considering that the information from intrusive and NILM enables the detection of exact appliance usage within a household [141].

A number of different disaggregation methods exists for the implementation of NILM, based on a general premise that can be summarised in three steps: 1) a centralised sensor at the main circuit-breaker is installed for data acquisition (usually recording active power, reactive power, voltage and current – but this depends on the disaggregation approach), 2) a library of specified features or signatures corresponding to each appliance is selected and 3) an algorithm is used, able to extract the features and accordingly disaggregate the load [142]. Recent developments allow steps 2 & 3 to run in conjunction and newly installed metering devices can identify new appliances that are used by the customers. The disaggregation algorithms depend on the electrical parameters recorded and the measuring frequency. Lower time-resolution (1 min. – 1 Hz) is usually associated with feature extraction based on steady-state step changes (i.e. event detection) while higher frequency measurements (1 Hz to 1 MHz) allow for harmonic analysis and identification of both transients and noise [143]. Computationally, disaggregation has been implemented based on a number of methods such as: statistical analysis [144], rule-based pattern recognition and expert systems [145], support vector machines [146], artificial neural networks (ANN) [147], probabilistic approaches [148] and others. The success of disaggregation algorithms is usually evaluated experimentally, against actual data from appliance monitoring, or through the use of publicly available datasets [149]. The prolific research output of academia and industry has resulted in commercially available smart-monitoring devices that can estimate cost of appliance usage and even warn for the possibility of imminent faults. Such devices are however still expensive and electricity energy suppliers usually provide their customers with simpler smart-metering equipment, that typically records only low-resolution aggregate consumption patterns.

Despite the advances in LV-metering and disaggregation, MV-network operators still rely on substation data for the purposes of operation and control, for future planning considerations and for centralised management of supply-demand interventions, such as the scheduling of dynamically-teleswitched-meters (DTMs) [150]. Data privacy and data ownership issues restrict network operators' access to high-quality end-user data and the only available option is often limited to the results of survey-based analysis. In particular, the government regulator for gas and electricity markets (Ofgem) in the UK, sets strict criteria for the use of SM data by DNOs [151] and a recent open letter to DNOs and other interested parties (Sept. 2016, available in [152]) indicates that the privacy-concerns are not resolved and that acquiring aggregated SM data is not an easy, straightforward process. Due to these limitations, disaggregation at the MV-level can offer better understanding of the demand characteristics in the absence of SM data and together with temporal decomposition and customer-classification, it can be considered part of a more detailed and comprehensive demand profiling. The results

of MV-disaggregation can also be used for studies of bottom-up composite load-models, which often rely on customer-surveys and nation-wide statistics for validation, as in [76]. However, and unlike SM-data, approaches for the disaggregation of MV-demands cannot always rely on high resolution measurements, as recorded parameters are often (but not always) limited to active power and reactive power of an hourly or half-hour resolution.

Authors in [153] present a model for the disaggregation of total demand into weather-related (temperature and humidity) and illumination-related loads, for the residential sector. However, the model does not take into account the multicollinearity effects between temperature and solar irradiance (correlations between the two predictors), as well as the possibility that loads which are not regarded as weather-related (such as cold loads or the use of electronics) might actually have seasonal variability. The authors also present a detailed disaggregation into household appliance usage, modelled by an agent-based analytical tool and relying on appliance-cycles (power-use) and time of use surveys for the model development. Authors in [154] present a multivariate polynomial model for determining load composition at the distribution level. Their approach is based on load-signatures for typical load-categories, but their use of harmonic analysis requires high resolution current-waveforms. Similarly, authors in [155] propose the use of ANN for determining load composition by using harmonic characteristics of typical load categories as inputs to the model. A detailed and comprehensive approach to MV-load disaggregation is proposed in [156] and [157]. The authors present an ANN-based method aimed at real-time load disaggregation. Datasets comprised of active power, reactive power and voltage are generated using Monte-Carlo simulation and used for ANN-training and validation. Eight load-categories are represented by static, voltage-dependent load-models and all possible combinations of contributions from these categories are considered and expressed by appropriate weighting factors. Voltage and total loads are used as the ANN training inputs and the targets are the simulated weighting factors. The validation is based on the error between the simulated and ANN-estimated weighting factors, as well as between the estimated aggregate load (from the composite parts) and the aggregate load from the validation dataset. The authors further expand their methodology in [158] and combine load disaggregation and load forecasting, that enables day-ahead prediction of total load and load composition, validated against measurement data.

In this thesis two disaggregation approaches are presented based on the seasonal correlations between electricity demand and meteorological/analemma parameters, as well as on the active/reactive power relationship in the form of multiple-regression analysis and power factor transformations, on a per half-hour of the day basis, as described in Chapter 6. These methods

aim to combine the effects of external conditions with the electrical-load characteristics in a statistical context, without relying on the instantaneous *P-Q-V* relationships, as expressed in load-modelling studies and used in previous disaggregation approaches. The disaggregated demands represent seasonal thermal-resistive loads, seasonal non-thermal loads associated with changes in the use of appliances related to occupancy-levels (e.g. electronic devices), cooling-loads and lighting loads. For the purposes of this analysis, it was also necessary to determine the base (threshold or balance-point) temperature and solar irradiance levels in order to determine the points of commencement (i.e. periods of switching) of electrical heating/cooling and lighting devices.

The demand for heating and cooling loads is usually assessed through the use of heating-degree-days (HDD) and cooling-degree-days (CDD), as described in [159]. The corresponding temperatures for HDD and CDD are, however, frequently selected arbitrarily and do not always correspond to the specific locations and consumption characteristics at the target study area. The disaggregation approach in this thesis does not rely on HDD and CDD and the temperature and solar irradiance threshold values are calculated directly from the weather and demand datasets, corresponding to the locations studied. Similar approaches for determining the base values have also been used in [121] and [153], for purposes of demand forecasting and disaggregation, but concentrated on state-wide and city-wide aggregate demands, in Australia and the US, respectively and in both cases indicating the significant presence of air-conditioning loads. The current thesis presents the base temperature values, as estimated for the temperate (and temperate-maritime) climates of the locations corresponding to the available MV-substation datasets. More detailed discussions regarding the estimation and use of the base temperature/solar irradiance values, as well as for the disaggregation methodologies are provided in Chapter 6.

2.5 Load Forecasting

Load forecasting is recognised as one of the most important aspects of electrical power research and thousands of scientific papers have been published over the years. Forecasting approaches vary according to the forecasting period, level of load aggregation, data availability, location, load characteristics and other parameters. Methods can be broadly categorised as time-series and causal (inverse-problem formulation), although hybrid models do exist, as in [160]. During the last decades, improved computational capabilities have also enabled the use of machine-learning approaches and ANN, e.g. [161].

Time-series methods aim to model the trend, seasonality and stationary processes of historical data by means such as: exponential smoothing, autoregressive moving-average (ARMA), autoregressive integrated moving-average (ARIMA), Box-Jenkins, and others. The analysis of dependencies is usually associated with causal methods, as the inverse-problem formulation is used to identify the main drivers of demand variability. Once the relationships have been established, future values are determined according to the variability of the independent variables (i.e. predictors). These approaches are, therefore, mostly formulated as regression problems, linear or polynomial, and taking into account the effects of one or multiple independent variables, as in [162]. As it is recognised that the electricity demand dependencies are often non-linear and multi-parametric, it is common practice to decompose demands according to hours of the day, days of the week, etc. in order to produce smaller subsets for which the dependencies can be linearly expressed, as discussed in [163]. Reviews and comparisons on available methods can be found in [164], [165], [166] and [167].

Due to the multi-parametric interactions between electricity demand and social, economic, weather and other components, there is no single method which can be considered as optimal or which can be shown to consistently perform well (and better than others) under all possible conditions. Similarly, performance comparisons are difficult due to the different scope of each individual study, although successful models' performances are, generally, considered to be below 10 % for long and medium term forecasts and below 5 % for short and very short term forecasts (in terms of the mean-absolute-percentage-error MAPE).

In Chapter 7, a multiple regression forecasting model is presented, for medium-term and short-term analysis. Model specifications are selected according to model-performances prior to the forecasting phase and a range of different combinations of predictors is considered. The parameter estimation is not restricted to linear relationships but allows for 2nd and 3rd degree polynomial expressions. The forecasting performances are also evaluated for models including the analemma variables only, which can be considered useful for load forecasting in the absence of meteorological data, or when reliable weather-forecasts are unavailable. Results are presented for both active power and reactive power forecasting.

2.6 Statistical Approaches

Chapters 3-7 provide discussions on the methodological challenges and appropriate statistical approaches, based on the experiences gained through the course of this PhD study. Where possible, the strengths and limitations of these approaches are also demonstrated and compared with examples.

Due to the diversity of applications of data analysis in electrical power studies, there is no generally agreed framework regarding preliminary data processing (or data pre-processing). Data normalisation/standardisation is discussed in Chapter 3, where the necessity for this pre-processing arises from the fact that the large set of available data corresponds to MV-substations with different levels of absolute demands. Previous studies have presented results in units of actual power and/or energy, e.g. [106], because the authors dealt with consumptions at comparable levels (e.g. from households with similar power requirements). In other cases, per-unit normalisation (i.e. with respect to maximum values) is used, as in [98] and [111], where the authors present consumption trends, averaged over customers with various power requirements; or as in [168], where voltage is normalised with respect to nominal values. In [124], normalisation is performed with respect to the mean values, while other normalisation methods such as min-max, z-score, etc. are discussed in [169], [170] and are also used in this study.

In this thesis, normalisation with respect to percentiles (e.g. actual values divided by the 95^{th} percentile value) is also explored. This approach reduces the effects of outliers as it is, in many cases, impractical to go over every single data-point that largely deviates from the mean and manually inspect whether it is an actual outlier. In many cases such values are valid observations and therefore filtering would cause unnecessary reduction in granularity. Nevertheless, other authors present justifications for the use of filtering as a means of eliminating outliers, as in [82] and [171]. These studies are however, concentrated on the pre-processing of data indented to be used for static and dynamic load modelling, where such extreme data-points affect the analytical calculations. Conversely, this thesis presents pre-processing approaches within the frame of statistical analysis where, given the large dataset used, inconsistencies are "averaged-out". It should be noted however, that the MV-data used comes from a larger dataset and that the selection process effectively discarded substation-data with error-measurements, missing values or large portions of outliers.

Normalisation is also used as a way of "highlighting" specific features of demand patterns that can be used for further analysis (e.g. Chapters 3 and 5). While processing prior to feature extraction is used by other authors, e.g. [170], this thesis presents different combinations of normalisation approaches and weighting factors, demonstrating their applicability specifically for the analysis of data from MV-substations and does not rely on a single approach.

Data filtering prior to regression analysis is also discussed. Various filtering approaches are used in literature such as: moving-average, Kalman and Savitzky-Golay filters, local-regression and others, particularly in the context of forecasting or, as mentioned before, for

reducing the effects of outliers. Examples of the use of such filters are found in [171], [172], [173] and [174]. Justifications for not using filtered values as inputs in the correlation and regression analysis, as well as examples of cases where "smoothed" variables are appropriate, are given in Chapter 4. Other issues related to the analysis of demand dependencies, such as non-linearity, multicollinearity, and the appropriate window-length for a moving-window regression approach, are discussed within the context of the corresponding methodologies, in Sections 4.3, as well as in Chapters 6 and 7.

Chapter 3: Profiling of Aggregate Demands

Profiling is defined, in the context of this thesis, as the graphical and numerical representations of the actual or normalised magnitudes of active power – *P*, reactive power – *Q* and voltage – *V*, with respect to different temporal-scales and statistical metrics. The purpose of this chapter is threefold. Firstly, to introduce the available datasets, as recorded at the MV-levels of distribution networks; secondly, to determine their main temporal-modes of variability; and thirdly, to profile the parameters, according to the most significant of these temporal-modes. In each stage, discussions are provided regarding the use of the results in the subsequent chapters and therefore, the analysis is also indented to lay the foundations for the work presented throughout this thesis.

Data pre-processing and the use of different normalisation and standardisation approaches is discussed in Section 3.2. No prior assumptions are made about the *P*, *Q* and *V* variabilities and these are investigated using Fourier decomposition, in Section 3.3. The diversity in the significance of hourly, weekly, yearly and other cycles, for different GSPs, is demonstrated, as well as the possibility of grouping according to these distinctions and the average contributions of each cycle to the total range of variations is presented.

Sections 3.4, 3.5 and 3.6 present profiles of weekly, hourly and yearly cycles. The differences between the normalised values within each time-scale are quantified and analysis-of-variance is used to determine the periods of similar and dissimilar levels, for the measures of central tendency. Section 3.7 presents the probability of occurrence of maximum (peak) and minimum demands, with respect to the diurnal as well as the seasonal perspectives and characteristic patterns related to customer-classes and external variables are illustrated. In Section 3.8, profiling of the rate of change of active/reactive power is presented and discussion is provided for the resulting patterns, with respect to thermal heating and lighting loads, as well as for the possibility of determining connections between these changes and the occurrence of faults and long/short interruptions in the distribution networks.

3.1 Description of Available Demand Data

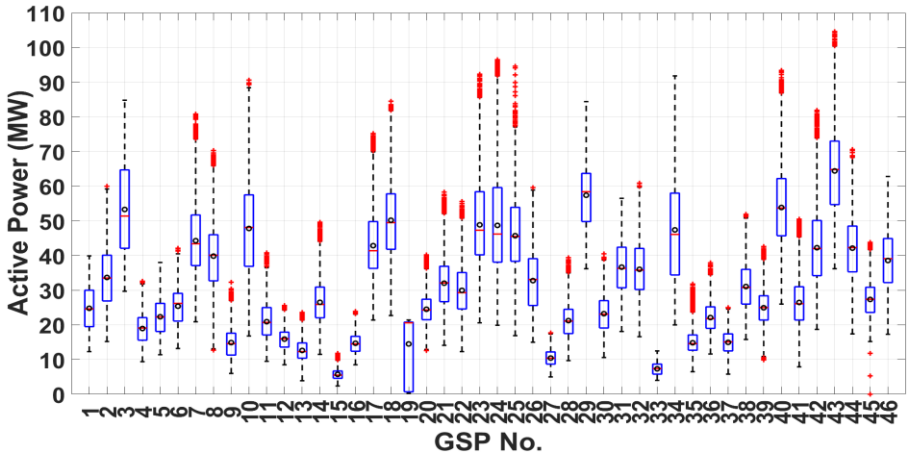
Available measurements from 98 medium-voltage (MV) grid-supply points (GSPs) were used for the analysis and results presented in this thesis. The datasets were obtained from primary and secondary distribution substations at four different geographical locations, i.e. Scotland, England, Denmark and Slovenia and are summarised in Table 3.1. Sampling rates are at a 30-minute resolution for Scottish-A, Scottish-B and Slovenian GSPs, adjusted to a 30-minute

resolution for the English GSPs (originally at 1s resolution) and at a 60-minute resolution for the Danish GSPs. The duration of measurement period is one year for all datasets, except for Scottish-B, for which six years of measurements were available. The corresponding sampled variables are: active power (for all GSPs), reactive power (for Scottish and English GSPs) and voltage (for English GSPs), in units of: MW, MVAr and kV, respectively.

Table 3.1: Summary of available measurements

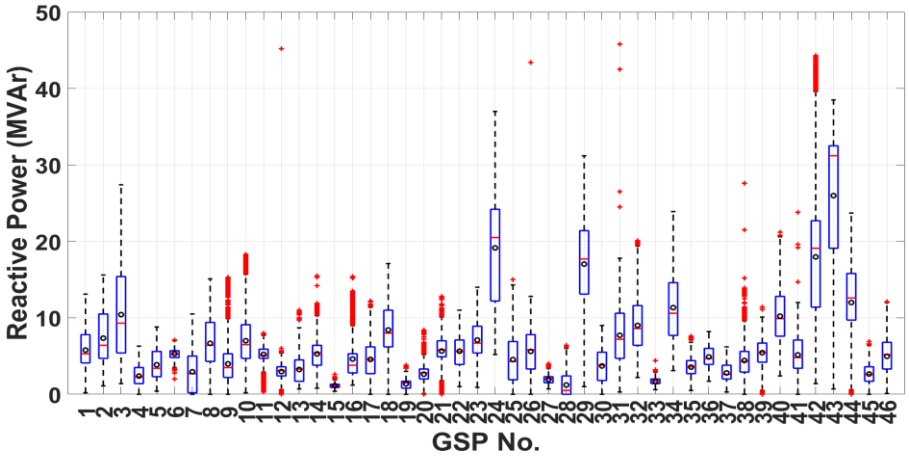
Name/Location	No. of GSPs	Variables	Duration	Voltage (kV)	Resolution
Scottish – A	46	P, Q	01/04/2009 – 31/03/2010	11/33	30 min.
Scottish – B	7	P, Q	01/01/2007 – 31/12/2012	11/33	30 min.
English	24	P, Q, V	01/01/2014 – 31/12/2014	6.6/11	30 min.
Danish	20	P	01/01/2013 – 31/12/2013	10	60 min.
Slovenian[6]	1	P	01/10/2007 – 30/09/2008	*[6]	30 min.

The basic statistical parameters for all 98 GSPs are presented in the form of whisker-box plots in Figures 3.1 (a-i), for each dataset and for each variable separately. The range of variations is represented as the extent of whiskers, while the extent of boxes shows 25th and 75th percentile values; lines in boxes indicate the median values (50th percentile) and circles in boxes indicate the mean values. Data-points shown as crosses are considered as outliers, based on their deviation from measures of central tendency (mean and median values) but are nevertheless, in most cases, valid observations and their extent can be justified by the characteristic demand patterns of the GSPs at which they were obtained (e.g. GSP-93 shown in Figure 3.7). The selected GSPs are part of a larger collection of available measurements, from which the presented datasets were selected based on the absence of actual outliers, faults in the measuring apparatus and/or missing values.

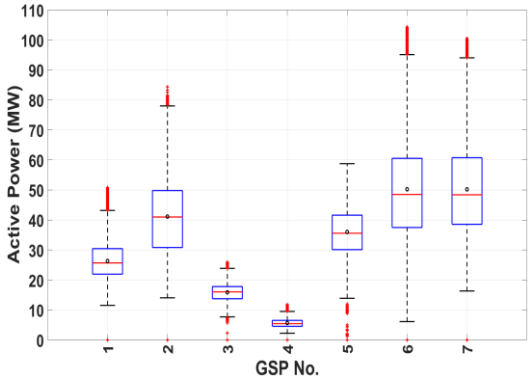


a) Active Power (Scottish-A)

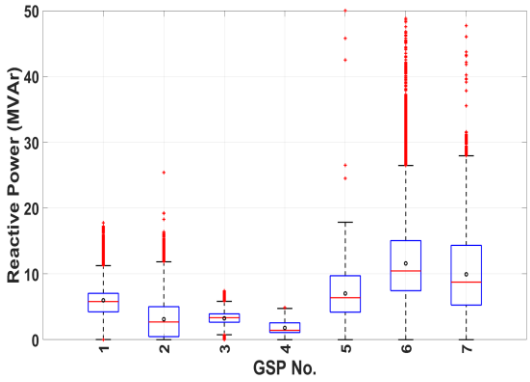
[6] The Slovenian data comes from SM-measurements of customers at the LV-level, corresponding to a single geographical location. These are aggregated and treated, for the purposes of this study, as data at the MV-distribution level.



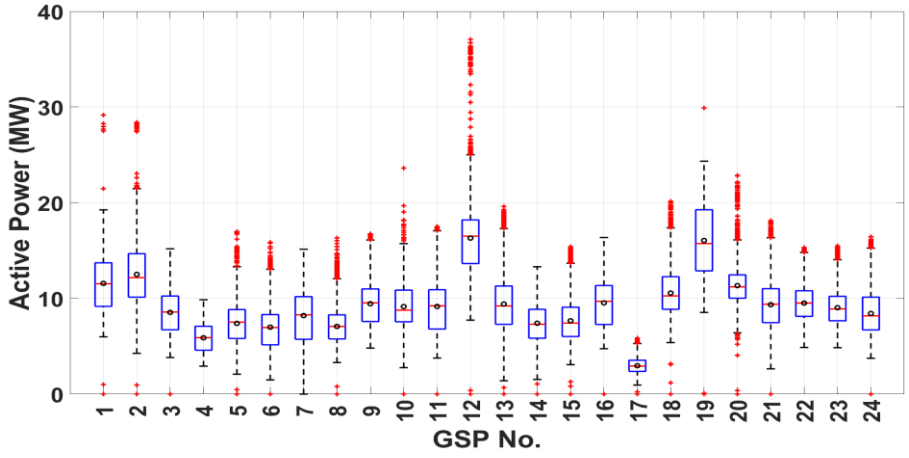
b) Reactive Power (Scottish-A)



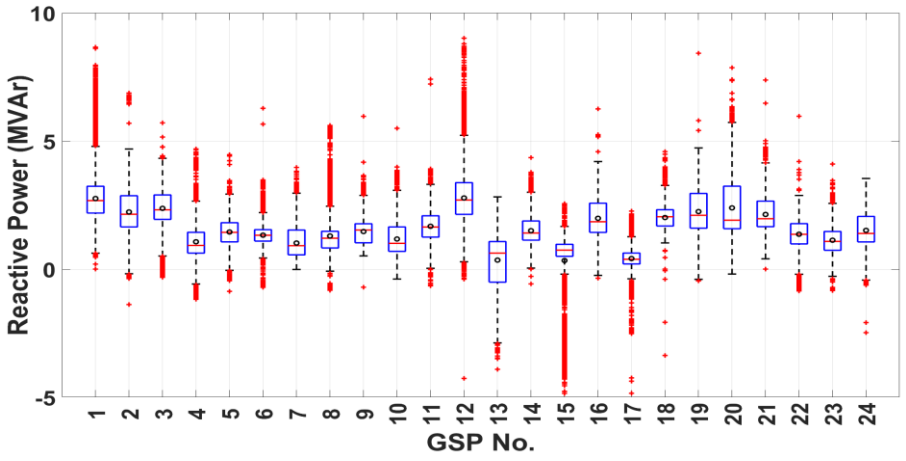
c) Active Power (Scottish-B)



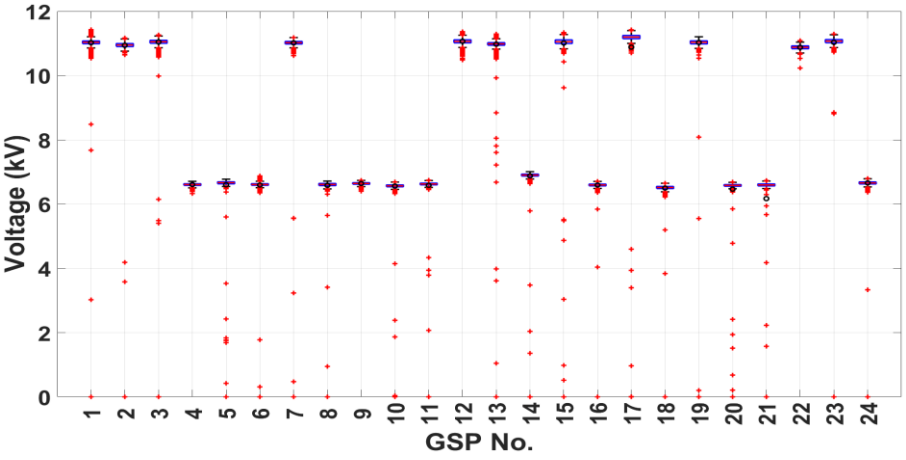
d) Reactive Power (Scottish-B)



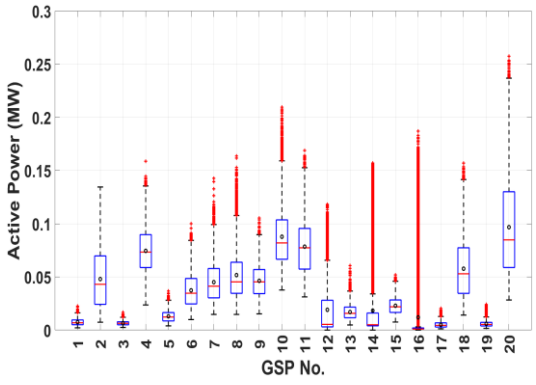
e) Active Power (English)



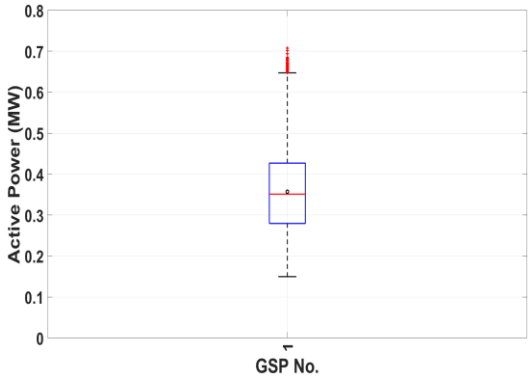
f) Reactive Power (English)



g) Voltage (English)



h) Active Power (Danish)



i) Active Power (Slovenian)

Figure 3.1: Basic descriptive statistics for the available measurements

3.2 Pre-processing: Normalisation, Standardisation and Scaling

Within the scope of preliminary data analysis (or data pre-processing), it is important to consider the conversion and adjustment of available datasets to common scales. This is necessary when examining samples from various populations, such as in the case of MV measurements, where the absolute values of measured active power, reactive power and voltage can be incomparable between GSPs due to operating conditions (e.g. measured *P* among GSPs vary in the MW range). The terms normalisation, standardisation and scaling can have different meanings depending on specific applications, but are also frequently used interchangeably. For this discussion, normalisation is not to be confused with data transformations, aimed at converting non-normal data to fit a Gaussian distribution.

One of the most commonly applied normalisation methods is the use of standardised values, also referred to as z-scores:

$$x_z = \frac{x_i - \bar{x}}{\sigma_x} \qquad (3.1)$$

where x_i is the i^{th} observation/value of variable *x*, $\bar{x}$ and σ_x are the mean and standard deviation values over all *x* observations and x_z is the resulting z-score value for the corresponding observation *i*. The values are thus adjusted to a scale representing their difference from the mean, in units of standard deviation. Although this can be convenient for comparisons between demands from various GSPs, the new scale can be misleading when dealing with samples which deviate from the normal distribution. Therefore, even though standardised values can be used for the presentation of demand datasets, care should be taken when using these values as inputs for statistical and analytical methods.

Another common approach is the scaling of measurements so that they are adjusted in a range [0,1], referred to as feature scaling, i.e.:

$$x_f = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad (3.2)$$

where each measurement, x_i, is rescaled by subtracting the minimum value of dataset *x*, i.e. x_{min} and expressing the result as a ratio to the range of measurements: $x_{max} - x_{min}$. Feature scaling produces comparable datasets with values restricted within the preassigned range, which is not necessarily [0,1]. However, under certain conditions this approach can also be misleading. Consider, for example, datasets with maximum or minimum values that largely deviate from the measures of central tendency (such as mean or median values) but are, nevertheless, "valid" observations (cannot be considered as outliers). In such cases the rescaled values will cluster around the lower/upper bounds of the new range, which will lead to false

conclusions when comparing with another dataset that does not have "extreme-value" observations. Similar issues can arise when rescaling measurements as ratios to maximum or mean values, such as:

$$x_j = \frac{x_i}{x_{(max,mean)}} \tag{3.3}$$

Possible solutions include: the normalization with respect to the Nth (e.g. 95th) percentile maximum value (lower sensitivity to outliers), although this will result in rescaled demand levels over unity (for measurements over the prescribed percentile) and scaling with respect to the median value (50th percentile), which is less sensitive to the skewness of distributions. Normalisation by the removal of the mean (or median) is also considered and relates to the z-score approach, but without scaling over standard deviation and allows for comparisons between the range of variations of datasets of different absolute demands.

$$x_k = x_i - x_{(mean,median)} \tag{3.4}$$

Further complications regarding data normalisation arise when presenting or analysing data of multiple dimensions. The number of possible normalisations rapidly increases with increased data dimensionality, as well as with respect to the desired relationships between rescaled data-points, which are determined by the analysis objectives. For example, when considering arrays of raw active power demands from a single GSP at a 30-minute resolution (i.e. $365\ \text{days} \times 48\ \text{half hours} = 17520$ data-points), data can be normalized with respect to the absolute maximum demand, 95th percentile maximum demand, mean demand, using standardised values (z-scores), by subtracting the mean value, etc. However, when dimensionality increases, such as when we want to examine demand variations in daily/diurnal profiles (demands per half-hour of the day in a diurnal perspective), normalisation is now possible with respect to global statistics, e.g. overall maximum demand, but also with respect to the newly available dimensions: 1) normalisation of demands per half-hour of the day with respect to the maximum value at each half-hour throughout the year, or 2) normalisation for each day of the year over the 48 half-hours of the day; and for all possible normalisation methods. When scaling the diurnal demands over the yearly/seasonal dimension, the normalised values capture deviations from the mean that can be useful when considering demand levels at particular half-hours of the day, compared with the rest of the year, e.g. for examining levels of heating/cooling loads. Similarly, when normalising with respect to each diurnal cycle, the extend of the variability between days of different absolute demands becomes comparable. Thus, since normalisation essentially requires operations between data-points and statistical parameters, the selected methods and combinations of methods can

"highlight" different characteristics and reveal patterns in the data that can be useful for different applications.

The various normalisation methods used in this thesis are explicitly identified and their use is justified, including cases when the results are validated with more than one approach. In particular, normalisations according to (3.1) and (3.3) are most frequently used. Furthermore, as discussed in Chapter 1, Section 1.4, the terms "daily" and "diurnal" are used interchangeably, as well as the terms "yearly", "annual" and "seasonal", to denote variations with respect to the daily cycle (00:00 to 23:30 hours) and the yearly cycle (January to December).

3.3 Temporal Decomposition and Periodicities

The variabilities of measured electrical parameters (i.e. *P*, *Q* and *V*) in the time-evolution sequences are, essentially, functions of both deterministic and stochastic processes, in the sense that it is practically impossible to predict the exact demand levels at any arbitrarily chosen future time. However, and assuming that electricity consumers are rational agents in need of electrical power based on external parameters and socio-behavioural patterns, the deterministic components have to be the most prevalent of the two, thus the success of forecasting models in predicting future demand levels within acceptable error boundaries. The dependencies of electricity demand to exogenous variables (such as weather) are discussed in Chapter 4 while forecasting models are discussed in Chapter 7. It is, however, necessary to initially determine the inherent characteristics of the available measurements and specifically the most important oscillatory temporal-modes of variability.

For the purpose of identifying periodicities in measured demand data, discrete Fourier transforms (DFT) are considered and implemented using a fast Fourier transform (FFT) algorithm [175], applied to vectors of measurements *X*:

$$Y(k) = \sum_{j=1}^{n} X(j)\, W_n^{(j-1)(k-1)} \tag{3.5}$$

$$W_n = e^{(-2\pi i)/n} \tag{3.6}$$

where W_n is the n^{th} root of unity, *j* and *k* are indices from 0 to $n-1$ for a vector of length n, *i* is the imaginary unit and *X* takes the values of active power or reactive power or voltage, for the corresponding datasets described in Section 3.1. For each GSP for which these variables are available, *Y* are the resulting complex-valued data in the frequency domain.

An example of active power demand decomposition into frequency components is shown in Figure. 3.2, for GSP-57. Figure 3.2 (a) shows the original active power demand measurements

at a 30-min. resolution for the duration of one-year (01/01/2014 – 31/12/2014) and Figure 3.2 (b) shows the normalised magnitudes at corresponding frequencies, which are the real-parts of complex vectors Y, normalised over the vector's length (i.e. 17520 for half-hourly data and 8760 for hourly data).

The highest frequency components in Figure 3.2 (b), for this particular GSP are found at: 1, 52, 104, 365 and 730 cycles/year and are referred to as: yearly (or seasonal), weekly, half-weekly, daily (or diurnal) and half-daily oscillations, respectively. It should be noted, that the purpose of the analysis is to determine the most significant frequencies in terms of normalised magnitudes, i.e. the periodicities with which P, Q an V vary throughout the course of one calendar year. In this context, a more generalised form of harmonic analysis is outside the scope of the study and similarly, the discussion provided does not go into spectral-leakage considerations.

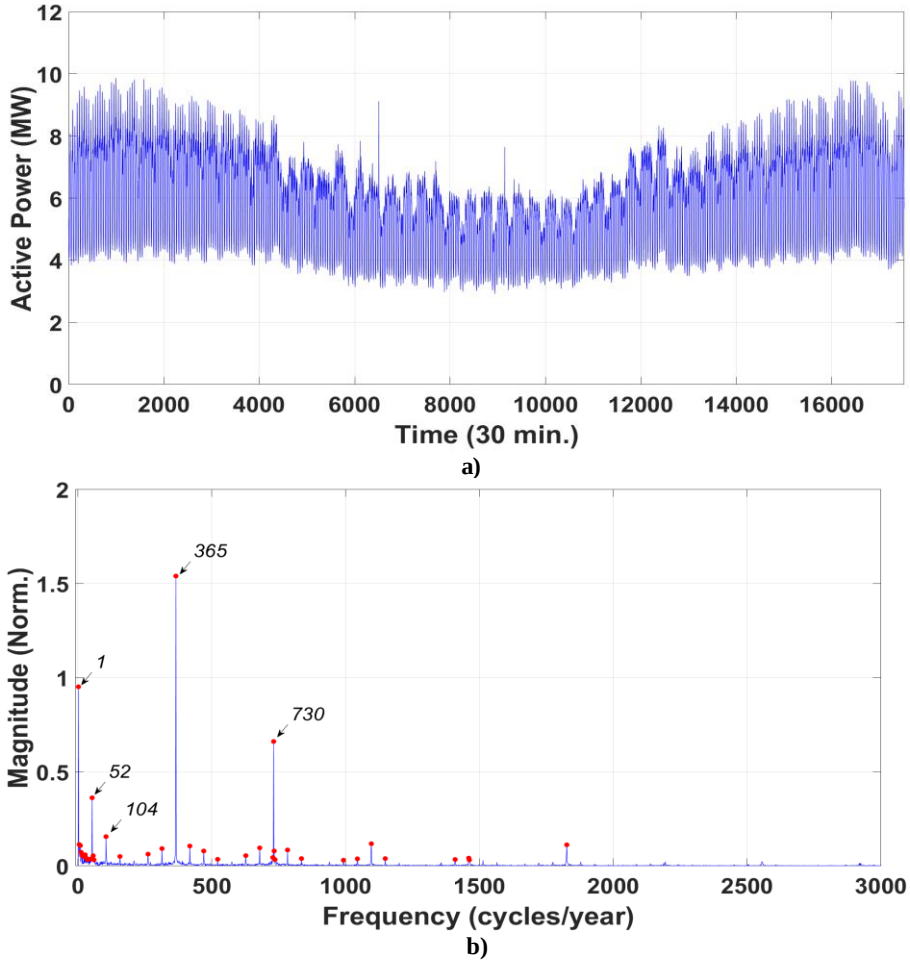


Figure 3.2: Example of FFT results for GSP-57 for active power: a) original signal, b) normalised magnitudes at corresponding frequencies

The first four components, for active power (GSP-57), are presented in detail in Figure 3.3. These are ordered from higher-to-lower normalised magnitudes and correspond to: the daily

variations in (a), the seasonal variations in (b), the half-daily variations in (c) and the weekly variations in (d). Note that the results for (a) and (c) are zoomed-in to show the cycles within a period of seven days.

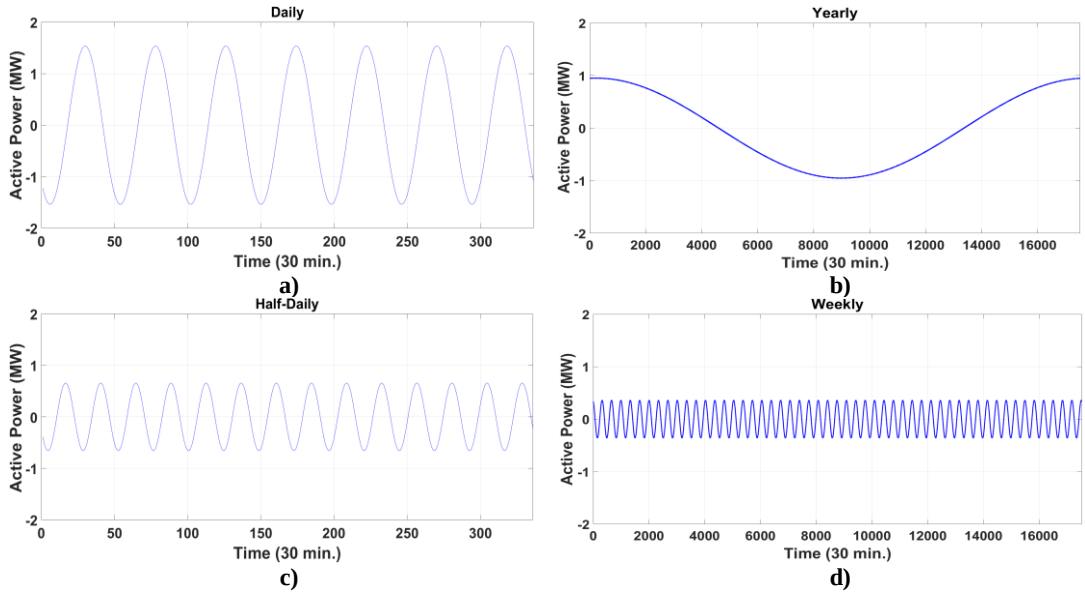


Figure 3.3: First four frequency components of FFT: a) daily, b) yearly, c) half-daily and d) weekly.

Figure 3.4 shows the original and "reconstructed" active power data (signal) for the same GSP, where the reconstructed data is based on the identified FFT frequencies with the ten highest magnitudes, plus the original trend-value (i.e. mean annual active power demand). The reconstructed time-domain data is obtained using an inverse fast Fourier transform (IFFT) of the complex vector Y which converts the frequency-domain components back to the time-domain X:

$$X(j) = \frac{1}{n}\sum_{k=1}^{n} Y(k) W_n^{-(j-1)(k-1)} \qquad (3.7)$$

where the terms are as defined in (3.6).

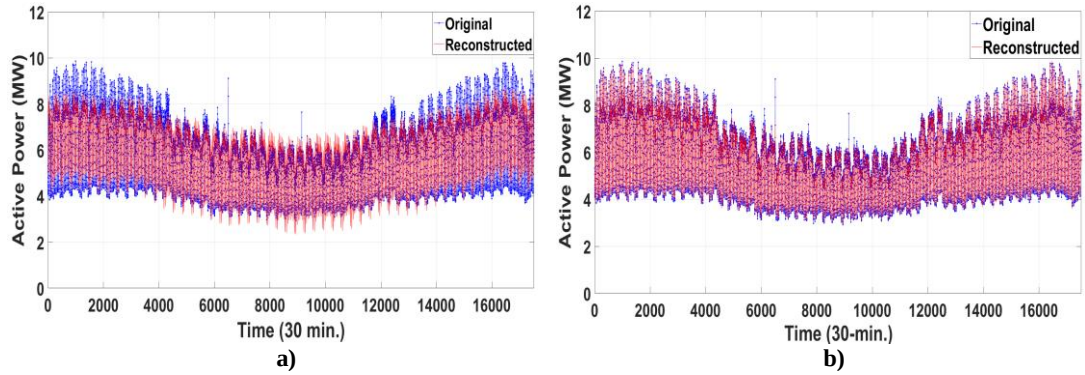


Figure 3.4: Original and reconstructed active power demand, for GSP-57, using: a) the first 10 FFT components and b) the first 1000 FFT components

3.3.1 Active Power

Figure 3.5 shows the probabilities for different frequencies being present in the first four components (i.e. with the highest magnitudes) for active power demand, based on the analysis which is now performed for all 98 GSPs. As mentioned, the most common frequencies correspond to daily, yearly, weekly and half-daily cycles.

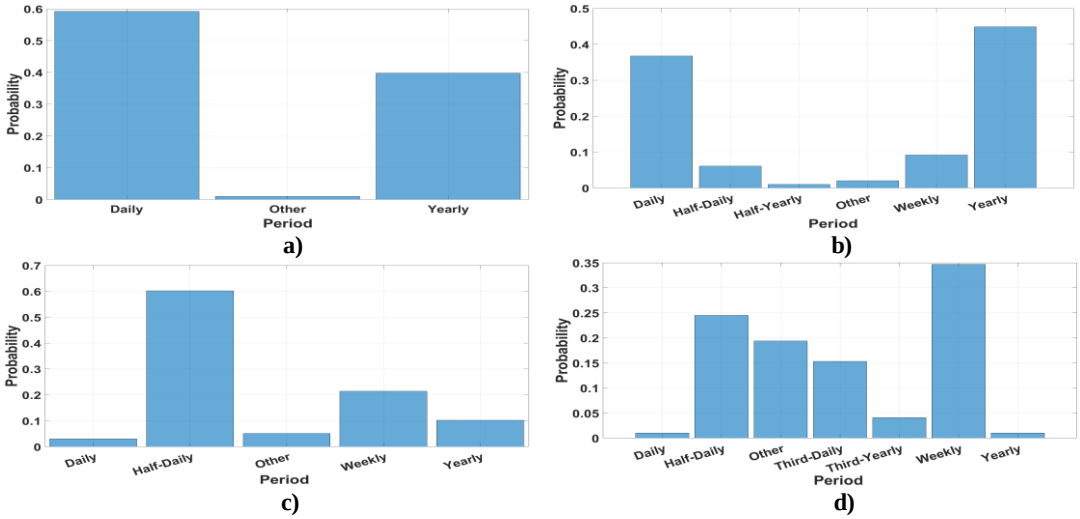


Figure 3.5: Probability of different cycles being present in the: a) first, b) second, c) third and d) fourth components for active power, for 98 GSPs

The relative contributions of these cycles to the overall active power demand variations are presented in Figure 3.6, where they are expressed as the normalised magnitudes using their ratio to the mean demand (per GSP). This now allows for comparisons between GSPs of different overall demand levels and for identification of inherently different types of GSPs.

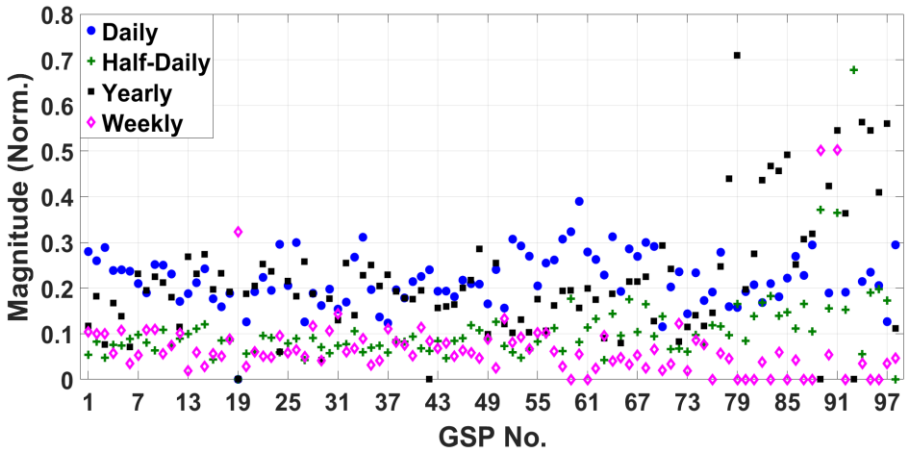


Figure 3.6: Normalised magnitudes (ratio to mean demand) of different frequency components for active power, for 98 GSPs

The differences shown in Figure 3.6 can be used to group GSPs according to the significance of different frequencies in their overall active power demand variability. This is demonstrated

in Table 3.2, where GSP-groups are presented according to the order of daily (D), yearly (Y), weekly (W) and half-daily (HD) cycles. For example, GSP-1 has the highest magnitude for the daily cycle, followed by yearly, weekly and half-daily cycles. It is therefore classified in Group-3 (Table 3.2), as a D, Y, W, HD GSP.

Table 3.2: GSP-groups according to the order of: daily (D), yearly (Y), weekly (W) and half-daily (HD) normalised magnitudes (for active power)

Group No.	Order of Norm. Magnitudes	GSP No. (out of 98)	Occurrence
1	Y, D, HD, W	7, 13, 14, 15, 17, 20, 21, 22, 23, 25, 32, 35, 36, 47, 48, 50, 71, 78, 80, 81, 82, 83, 84, 85, 87, 88, 90, 92, 94, 95, 96	31.63 %
2	D, Y, HD, W	4, 10, 26, 33, 38, 39, 40, 43, 45, 46, 49, 57, 58, 59, 60, 61, 62, 64, 66, 67, 68, 69, 73, 74, 75, 76, 77, 86	28.57 %
3	D, Y, W, HD	1, 2, 5, 9, 11, 12, 28, 30, 34, 41, 44, 52, 53, 54, 55, 56, 91, 98	18.36 %
4	D, W, Y, HD	3, 24, 31, 51, 63, 72	6.12 %
5	Y, D, W, HD	8, 16, 18, 27, 37	5.10 %
6	Y, HD, D, W	70, 79, 97	3.06 %
7	D, HD, Y, W	6, 29, 65	3.06 %
8	D, W, HD, Y	42, 89, 93	3.06 %
9	W, D, HD, Y	19	1.02 %

Almost all GSPs have either daily (~58 %) or yearly (~ 40 %) cycles as the predominant modes of variability and ~80 % of the GSPs have daily and yearly cycles as one of the first two components. Differences in the ordering and magnitudes of the dominant frequencies can be used for purposes of dependence analysis, as Fourier transforms can show similar periodicities in the predictor variables; for load type disaggregation, as the higher yearly/seasonal components usually correspond to weather-affected loads (e.g. electrical heating or cooling loads) and for the classification of GSPs into corresponding customer-classes. The classification approach in this thesis (Chapter 5) does not explicitly use the Fourier components, however, based on the results of the current section, the seasonal and diurnal (daily) demand patterns are used, as it is indicated that these can provide distinctions among GSPs. Similarly, analysis in Chapters 4 and 6 is conducted on a daily, seasonal and seasonal per half-hour basis.

The results can also be used to identify GSPs with significantly different normalised magnitudes of dominant frequency components, which can be considered as "atypical" when compared to the vast majority of GSPs. An example is shown in Figure 3.7, corresponding to GSP-93 (i.e. GSP-16 of Danish dataset), which has daily variations close to 1.5 times the mean value and is, most probably, related to industrial consumption.

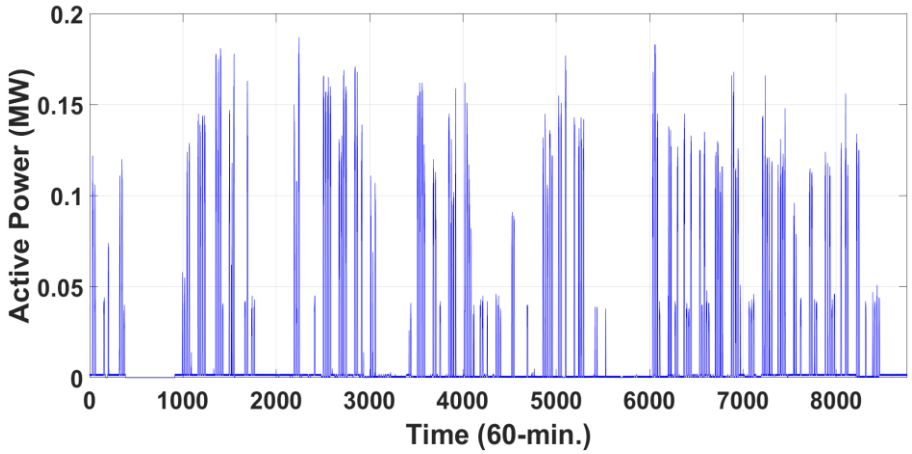


Figure 3.7: GSP-93 which shows "extreme" daily variations compared to the mean demand

Figure 3.8 shows the Pearson's correlation coefficient - *r*, between the original active power demand measurements and the signal reconstructed from the first ten components (with the highest magnitudes) of the FFT analysis. The Pearson's coefficient is defined in terms of the covariance of two variables *x* and *y* as:

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{3.8}$$

where *cov* is the covariance between *x* and *y* and σ_x, σ_y are their standard deviations. A more detail description and definitions of correlation coefficients is presented in Chapter 4.

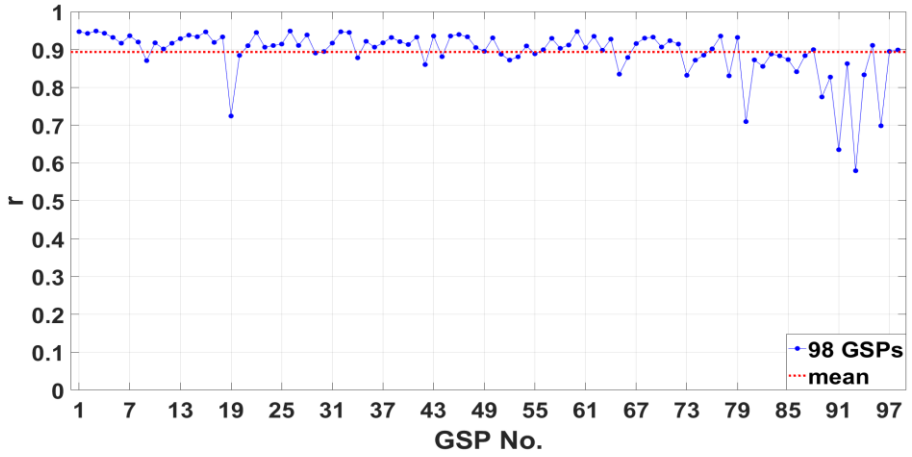


Figure 3.8: Correlations between the original and reconstructed signals for active power, for 98 GSPs

Although a stronger correlation does not necessarily imply that the original and reconstructed signals have the same absolute values, it is nevertheless a good indicator of whether the first ten components perform well in reconstructing the signal's variations. Furthermore, it can be seen that GSPs which are distinctively different in the results presented in Figure 3.6 also have

weaker correlations with the original signal, as shown in Figure 3.8. Apart from GSP-93, examples include: GSPs-91 and 19, which has zero normalised magnitudes for daily, half-daily and yearly cycles and non-zero normalised magnitudes only for weekly cycles (considering the first four components as presented in Figure 3.6 and Table 3.2).

The above results indicate that, generally, in the case of active power demand, a relatively small number of FFT components are able to reconstruct the original variations up to a very good degree, with an average correlation coefficient close to 0.9. This can be interpreted as lower penetration of "stochastic" active power demand variations, or in any case, reduced non-periodic variability of the signals. In the following sections (Sections 3.3.2 and 3.3.3), the results are presented for reactive power and voltage, based on the same analysis as for active power.

3.3.2 Reactive Power

Figure 3.9 shows the probabilities for different cycles being present in the first four components of the FFT analysis for reactive power demand and considering a smaller subset of 77 GSPs, for which reactive power measurements were available. These correspond to the first 77 GSPs used in the active power analysis (77 out of 98) in the previous section (for reference see Table 3.1).

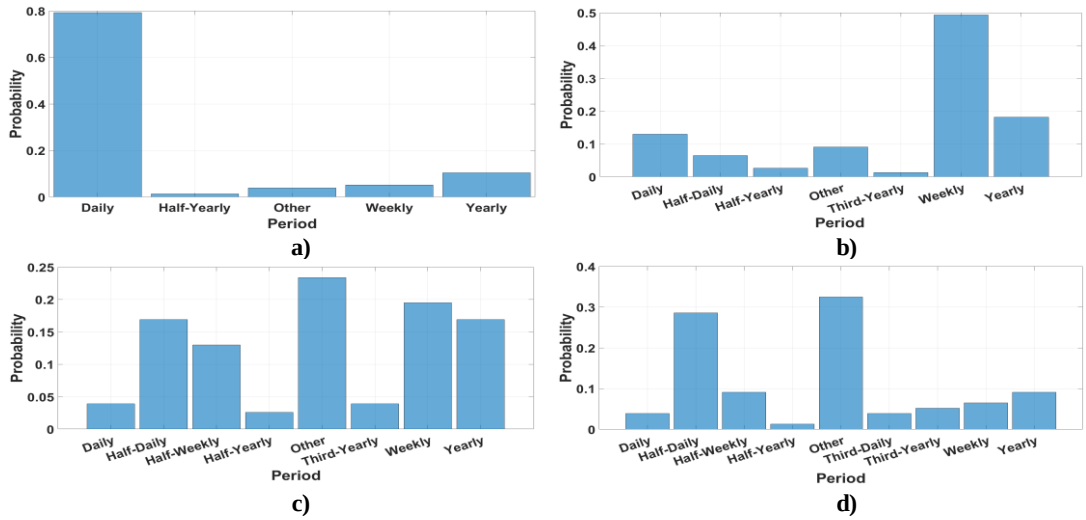


Figure 3.9: Probability of different cycles being present in the: a) first, b) second, c) third and d) fourth components for reactive power, for 77 GSPs

The daily cycles outweigh the importance of the yearly cycles (seasonality), indicating that variations of reactive power are more pronounced in the diurnal period. Weekly cycles are also present, primarily in the second component, while the third and fourth components are dominated by half-daily, half-weekly, weekly and yearly cycles. "Other" cycles are also present in the first four components and they correspond to frequencies of 313, 3, 7 and 4

cycles per year. Although these "other" periodicities have a high aggregate probability of occurrence, their individual probabilities are much lower and are not consistently present for all GSPs. In Figure 3.10, the normalised magnitudes of the daily, half-daily, yearly and weekly cycles are presented.

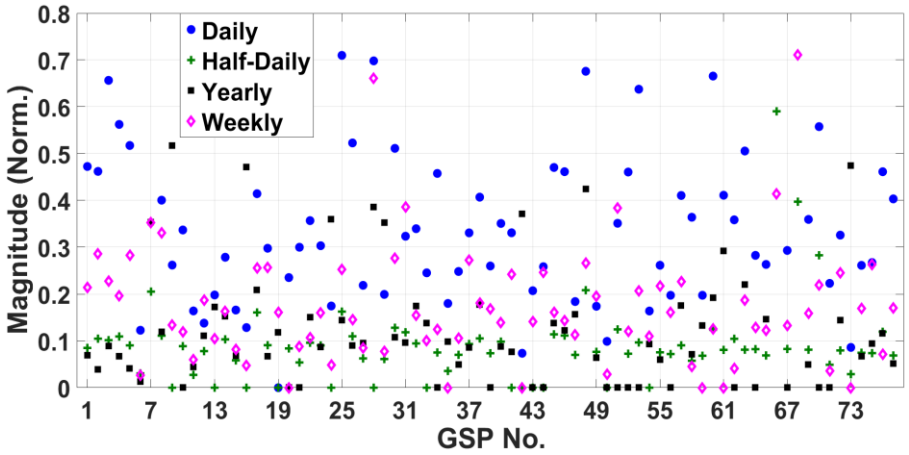


Figure 3.10: Normalised magnitudes (ratio to mean demand) of different frequency components for reactive power, for 77 GSPs

Similar to the analysis for active power, the normalised magnitudes can be used to determine GSPs that can be grouped together according to the ordering of these periods (excluding other periods shown in Figure 3.9). These are presented in Table 3.3. The vast majority of GSPs have daily cycles as the dominant (first) component (~80 %) and almost all of them have the daily component as one of the first two (~99 %). Seasonality (Y) is the dominant period for ~10 % of the GSPs and it is included in the first two components in ~30 % of all cases.

Table 3.3: GSP-groups according to the order of: daily (D), yearly (Y), weekly (W) and half-daily (HD) normalised magnitudes (for reactive power)

No.	Order of Norm. Magnitudes	GSP No. (out of 98)	Occurrence
1	D, W, HD, Y	1, 2, 3, 4, 5, 10, 18, 21, 23, 25, 26, 30, 34, 36, 37, 39, 40, 43, 44, 50, 52, 53, 55, 56, 64, 67, 69, 74, 77	37.66 %
2	D, W, Y, HD	8, 11, 14, 15, 17, 28, 38, 41, 45, 46, 54, 57, 72, 75	18.18 %
3	D, Y, W, HD	7, 13, 22, 27, 32, 33, 47, 48, 60, 63, 65	14.28 %
4	Y, D, W, HD	9, 16, 24, 29, 68	6.49 %
5	D, HD, W, Y	6, 62, 70, 71	5.19 %
6	D, Y, HD, W	35, 58, 59, 61	5.19 %
7	Y, D, HD, W	42, 66, 73	3.89 %
8	W, D, HD, Y	31, 49, 51	3.89 %
9	D, HD, Y, W	20, 76	2.59 %
10	W, D, Y, HD	12	1.12 %
11	W, Y, D, HD	19	1.12 %

As in the case of active power, the results can be used to identify GSPs with demands characteristically different than the rest of the set and to identify potential "outliers". As an example, GSP-68 (i.e. English GSP-15) has a normalised magnitude for the yearly cycle that is more than 2.5 times the mean reactive power demand. The reason becomes apparent when considering the original signal, shown in Figure 3.11. The "erratic" reactive power fluctuations between 8000-12000 half-hours, which also extent to negative values, increase the magnitude of the, otherwise, weak yearly component.

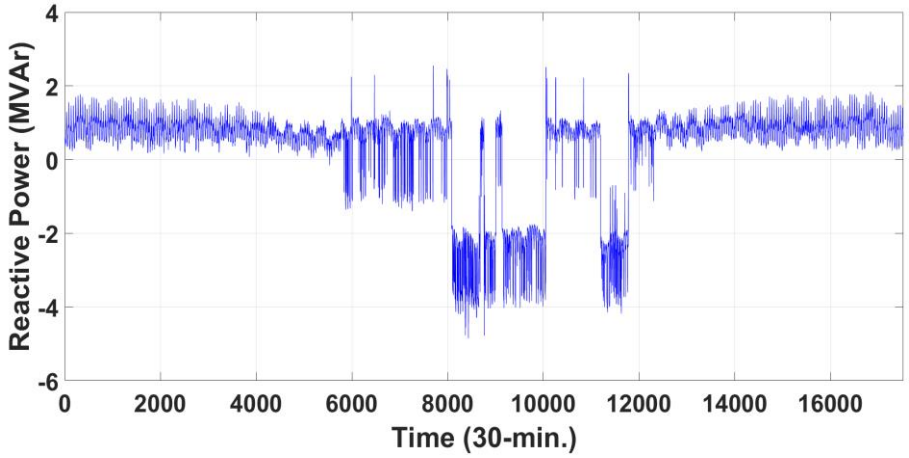


Figure 3.11: GSP-68 which shows ''extreme'' seasonal variations compared to the mean demand

Figure 3.12 shows the Pearson's correlation coefficients (3.8), between the original reactive power demand measurements and the signals reconstructed from the first ten components.

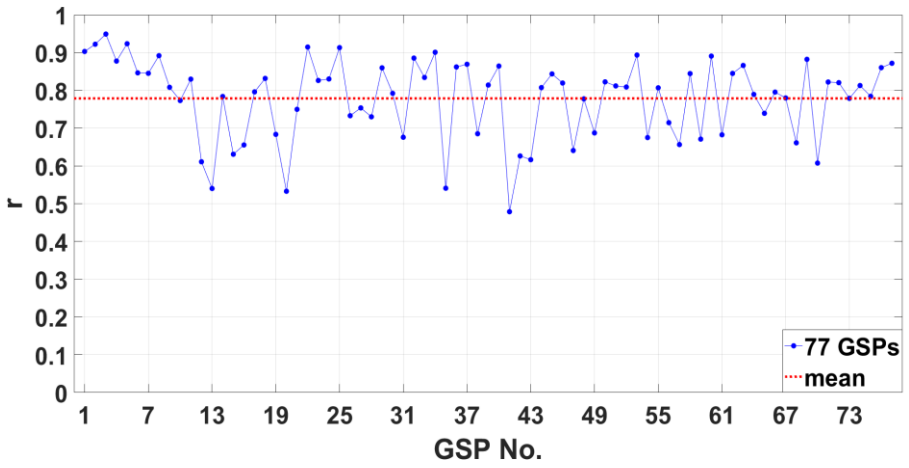


Figure 3.12: Correlations between original and reconstructed signals for reactive power demand, for 77 GSPs

The average correlation coefficient is lower than the one for active power (0.78 compared with 0.89) and there is also a higher variability of the results among GSPs, with correlation coefficients ranging from 0.95 to 0.45. This implies that reactive power includes significant

non-periodic components and/or stochastic variations compared with active power demands. This is further demonstrated in the analysis of correlations and dependences in Chapter 4 and in the load forecasting performances, discussed in Chapter 7.

3.3.3 Voltage

Figure 3.13 shows the probabilities for different cycles being present in the first four components of the FFT analysis for voltage and considering the 24 English GSPs, for which voltage measurements were available. These correspond to GSPs 54 to 77 out of the total of 98 and similarly to GSPs No. 54 to77, as presented in the FFT analysis of active and reactive power (for reference see Table 3.1).

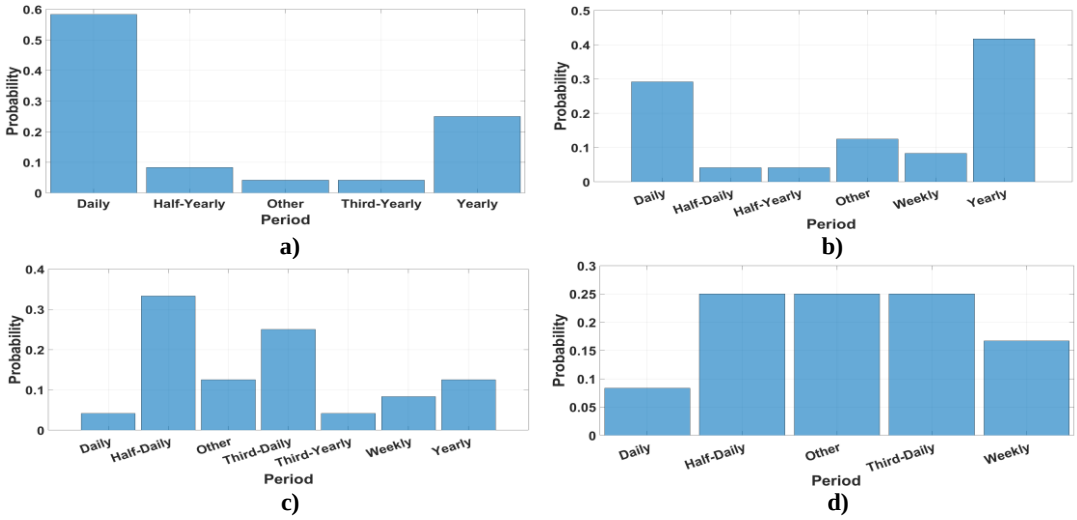


Figure 3.13: Probability of different cycles being present in the: a) first, b) second, c) third and d) fourth components for voltage, for 24 GSPs

Daily cycles are the most probable with respect to the first component, yearly and daily cycles considering the second component, while half-daily, third-daily and "other" cycles are shown in the third and fourth components. As the in the case of reactive power, the "other" component shows high probability, particularly in Figure 3.13 (d), but this is due to aggregation. The individual components constituting the "other" group have insignificant probabilities.

Figure 3.14 shows the normalised magnitudes of variations in the daily, half-daily, yearly and weekly cycles. These are at a scale of 10^{-3} indicating that the extent of variations is very low, compared to the corresponding variations for active and reactive power demands, which means that the values do not, on average, deviate significantly from the mean (also shown in the descriptive statistics in Section 3.1). This does not relate to the voltage scale being in kV, but rather to the allowed voltage fluctuations which must be, according to UK standards, within ± 6 % of nominal voltage levels (in MV distribution networks) [176].

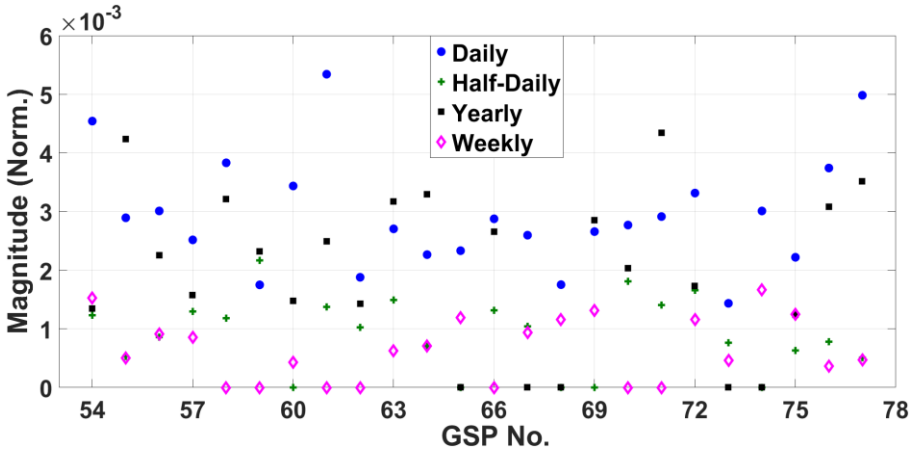


Figure 3.14: Normalised magnitudes (ratio to mean demand) of different frequency components for voltage, for 24 GSPs

Table 3.4 shows the resulting groups according to the ordering of normalised magnitudes and with respect to the daily (D), yearly (Y), weekly (W) and half-daily (HD) cycles. ~73 % of the GSPs have the daily cycle and ~24 % the yearly cycle as the dominant components. Even though the same characteristic frequencies (as in the case of active and reactive power) are present here as well, there are much weaker correlations between actual and reconstructed signals, based on the first 10 components, as shown in Figure 3.15.

Table 3.4: GSP-groups according to the order of: daily (D), yearly (Y), weekly (W) and half-daily (HD) normalised magnitudes (for voltage)

No.	Order of Norm. Magnitudes	GSP No. (out of 98)	Occurrence
1	D, Y, HD, W	57, 58, 61, 62, 66, 70, 72, 76, 77	37.50 %
2	D, W, Y, HD	54, 65, 68, 74	16.67 %
3	Y, D, HD, W	55, 63, 64, 71	16.67 %
4	D, Y, W, HD	56, 60, 75	12.5 %
5	D, HD, W, Y	67, 73	8.34 %
6	Y, HD, D, W	59	4.17 %
7	Y, D, W, HD	69	4.17 %

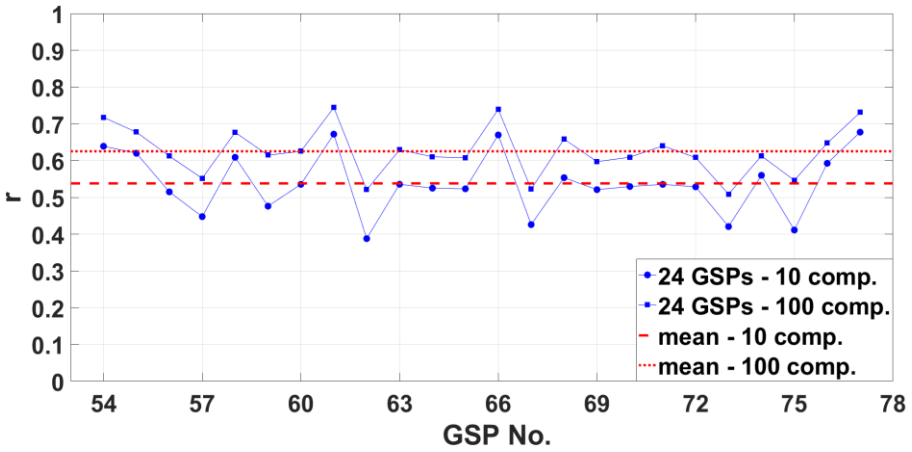


Figure 3.15: Correlations between original and reconstructed signals for voltage, for 24 GSPs

The correlation results can be improved (this applies to active and reactive power signals as well) with the inclusion of more components, as demonstrated in Figure 3.15 by the comparison of correlations between original voltage measurements and reconstructed data from the first 10 and first 100 components of the frequency domain.

Voltage levels are primarily determined by the operating configuration of the corresponding networks, at each level (e.g. at 6.6 kV and 11 kV for the above examples). The results indicate that the deviations from the nominal (which are generally below ± 6 %) cannot be successfully reconstructed from a limited number of FFT components which means that periodicities such as yearly or daily cycles are only weakly correlated with the voltage fluctuations. From a statistical perspective, the fluctuations can be considered as stochastic. This is further investigated in the profiling sections (3.4 to 3.6) and also in the analysis of dependencies in Chapter 4.

Table 3.5 summarises the results presented in this section, using the mean and standard deviation values (among all available GSPs) for the percentage-contribution of each component to the total range of variations (max. – min. values over a one-year period), for active power, reactive power and voltage. Since the results are averaged over all available GSPs, they are inappropriate for determining the characteristics of specific substations, which require individual analysis (presented in the previous sections) and thus the aggregate contributions from the various components can exceed 100 %, i.e. sum of values in columns of Table 3.5.

Table 3.5: Mean (and standard deviation) percentages to total range of variations for selected periodicities for: active power, reactive power and voltage

Periodicity:	% to Total Range of Variations		
	Active Power	***Reactive Power***	***Voltage***
Daily	41.43 (13.21)	36.79 (17.21)	15.21 (10.91)
Weekly	13.34 (7.94)	17.04 (10.23)	4.59 (3.88)
Yearly	33.23 (10.71)	13.67 (12.72)	14.81 (9.79)
Half-Daily	16.53 (4.75)	8.72 (4.69)	6.54 (4.17)
Third-Daily	5.45 (2.40)	3.36 (3.56)	6.70 (3.49)
Half-Weekly	4.83 (3.91)	6.05 (5.31)	0.60 (1.67)
Half-Yearly	0.35 (2.29)	1.63 (5.05)	2.65 (6.43)
Other	4.10 (1.19)	5.34 (1.43)	3.78 (0.82)

Integer-multiple frequencies corresponding to half-daily and third-daily cycles and similarly to half-weekly and half-yearly cycles, indicate intra-daily, -weekly and -yearly variabilities. For example, primarily residential GSPs have active power demand patterns with characteristic morning and evening peaks, separated by mid-day and mid-night troughs, resulting in stronger half-daily components, compared with primarily commercial GSPs (e.g. GSPs-14,15 and 52,53 in Figure 3.6). These differences in demand patterns are further

illustrated in the context of GSP-classification and customer-sector disaggregation in Chapter 5. However, it should be noted that the effects of interference (superposition) between these frequencies, have not been investigated any further and that in order to model realistic daily, weekly and yearly periodicities, the effects of the integer-multiple frequencies should be taken into consideration as well (this would also require individual GSPs analysis and the inclusion of the phase angles). It should also be noted that the periodicities presented in this section can be determined using other approaches such as autocorrelation analysis, which is related to Fourier transforms (Wiener-Khinchin theorem), by using wavelet-analysis or other spectral density estimations such as Welch's or Bartlett's methods and by periodograms [177], [178] and [179]. The following sections (3.4, 3.5 and 3.6) present profiling with respect to the most prevalent modes of variability, as determined through the results of the current section.

3.4 Weekly Profiling

The results presented in Section 3.3 indicate that there are substantial differences in demand levels for different days of the week, demonstrated by the presence of weekly components in the active power, reactive power and voltage variations, as deconstructed using the discreet Fourier transform analysis. To further investigate these differences, two methods are considered, the ANOVA test and the Kruskal-Wallis test, in order to determine whether there are statistically significant differences for each of the *P*, *Q*, *V* parameters among groups; where groups are defined as the different days of the week, i.e. *j=1,2, ... ,7* from Sunday to Saturday. The data preparation procedure is as follows:

- For each GSP, the values are standardised according to the z-score in (3.1), in order to produce data comparable among GSPs of different absolute demands.
- For each GSP, the mean value is calculated for each day of the week (i.e. for *j=1,2, ... ,7*).
- A 2-D array of the days of the week and all GSPs is created, i.e. 7×98 for active power, 7×77 for reactive power and 7×24 for voltage, or in general $j\times i$.

The Null hypothesis can be formulated as:

$$H_o: a_1 = a_2 = \cdots = a_j \qquad (3.9)$$

where *a* corresponds to each group's mean or median (according to the selected method). The first method is a one-way analysis of variance model (ANOVA), of the form:

$$y_{ij} = a_j + \varepsilon_{ij} \qquad (3.10)$$

where y_{ij} is the i^{th} observation or in this case the mean value for the i^{th} GSP for day j, a_j is the mean value for day j over all GSPs and ε_{ij} is the error. The total variation, or total sum of squares (SST) is partitioned into the intra-group variation or sum of square errors (SSE):

$$\sum_i \sum_j \left(y_{ij} - \bar{y_j}\right)^2 \qquad (3.11)$$

and the between-groups variation, or sum of squares due to group differences (SSG):

$$\sum_j n_j \left(y_j - \bar{y}\right)^2 \qquad (3.12)$$

where n_j is the sample size for each $j=1,2, \ldots ,7$, $\bar{y_j}$ is the mean value per day j and $\bar{y}$ is the mean value over all days of the week. The ratio of the SSG to the SSE (i.e. F-statistic) determines whether the Null hypothesis can or cannot be rejected, according to the selected p-value. The p-value can be defined as the probability of obtaining a result with a value which is greater than or equal to the observed value, when the Null hypothesis is true, purely by chance [180], [181], [182]. Therefore, small p-values indicate that it is unlikely that the observed effects (i.e. group differences) can be attributed to deviations of purely stochastic nature and, in the context of this study, they indicate different levels for the observed parameters *P*, *Q* and *V* between the given pairs (days of the week). The level of statistical significance is arbitrarily chosen and there is considerable controversy and often misuse regarding the presentation and interpretation of the p-values [183]. There is, however, a general "consensus" of using significance levels – a between 0.05 and 0.01, or 5% to 1% chance of rejecting the Null hypothesis when it is actually true (i.e. Type-I error, "false-positive"). The discussion provided for the results presented is based on a significance level $a = 0.05$.

The one-way ANOVA approach is generally referred to as the "means model". The model is based on assumptions of normality, independence and homoscedasticity of input samples and normality of the residuals between group observations and group means [184]. Although ANOVA models can function well under slight deviations from these assumptions [185] [186], a second, non-parametric approach is considered with no prior assumptions for the data distributions, in order to increase robustness and offer further validation for the results. This method is the non-parametric equivalent of the one-way ANOVA, namely the Kruskal-Wallis (K-W) one-way analysis of variance [187]. In the Kruskal-Wallis procedure, the analysis of variance is based on the ranks of input data and not the actual values and the Null hypothesis (3.9) is, in this case, defined in terms of medians not means (thus addressing issues of non-normality).

Figures 3.16, 3.17 and 3.18 show the resulting groups and Tables 3.3, 3.4 and 3.5 present the p-values for the Null hypothesis (3.9) (i.e. that there are no statistically significant group differences between the mean values for all possible pairs of the seven days of the week), for active power, reactive power and voltage, respectively. The p-values shown within brackets are the corresponding results for the Kruskal-Wallis test, for the modified Null hypothesis as applied to group median values.

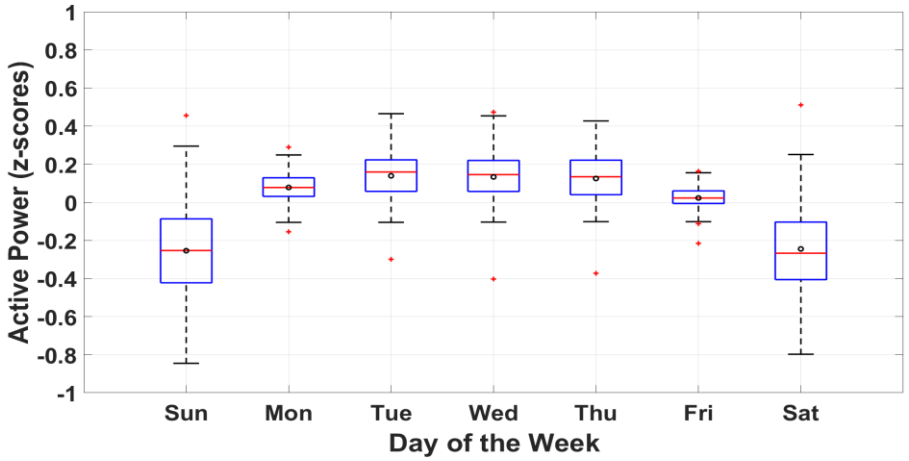


Figure 3.16: Groups of normalised active power for each day of the week based on 98 GSPs

Figure. 3.16 shows a clear, general tendency of higher levels of active power demand during the weekdays and particularly for Tuesdays, Wednesdays and Thursdays. Weekends are at a level of approximately -0.3 standard deviations from the mean. However, the extend of normalised demands is considerably higher for weekends, which indicates that there is higher load variability between the various GSPs for Saturdays and Sundays and also that the general tendency is probably violated for some GSPs. This is demonstrated and further discussed in Figures 3.19 and 3.20. The results for individual pairs of days are presented in Table 3.6.

Table 3.6: One-way analysis of variance (ANOVA) and Kruskal-Wallis tests results, for group differences between active power demand for the 7 days of the week

Pair:	p-value:	Pair:	p-value:	Pair:	p-value:
Sun – Mon	3.71×10^{-8} (3.71×10^{-8})	***Mon – Wed***	0.17 (0.15)	***Tue – Sat***	3.71×10^{-8} (3.71×10^{-8})
Sun – Tue	3.71×10^{-8} (3.71×10^{-8})	***Mon – Thu***	0.34 (0.31)	***Wed – Thu***	0.99 (0.99)
Sun – Wed	3.71×10^{-8} (3.71×10^{-8})	***Mon – Fri***	0.17 (0.02)	***Wed – Fri***	1.51×10^{-5} (3.19×10^{-7})
Sun – Thu	3.71×10^{-8} (3.71×10^{-8})	***Mon – Sat***	3.71×10^{-8} (3.71×10^{-8})	***Wed – Sat***	3.71×10^{-8} (3.71×10^{-8})
Sun – Fri	3.71×10^{-8} (3.71×10^{-8})	***Tue - Wed***	1 (1)	***Thu – Fri***	8.97×10^{-5} (2.03×10^{-6})
Sun - Sat	0.99 (1)	***Tue – Thu***	0.99 (1)	***Thu – Sat***	3.71×10^{-8} (3.71×10^{-8})
Mon – Tue	0.09 (0.06)	***Tue – Fri***	3.73×10^{-6} (3.71×10^{-8})	***Fri - Sat***	3.71×10^{-8} (1.41×10^{-7})

There are significant differences between mean (and median) active power levels for: Sundays with all weekdays and between Saturdays with all weekdays as well as between Fridays with Tuesdays, Fridays with Wednesdays and Fridays with Thursdays. The Null hypothesis cannot be rejected for: Sundays with Saturdays and all pairs of weekdays excluding the discrepancies mentioned for pairs with Fridays. There is also a very good agreement between the results obtained from the ANOVA and the K-W tests, except for the pair of Mondays-Fridays, for which the K-W test marginally rejects the Null hypothesis.

Figure 3.17 shows the resulting groups for reactive power. These are similar to active power, which is reasonable based on the underlying *P-Q* relationship and also from a statistical perspective, due to the strong positive correlations between the two, discussed in more detail in Chapter 4. Similar to active power, higher extend of the range is shown during weekends and group differences are apparent for Mondays and Fridays compared with the rest of the weekdays. These are also the days with relatively low extent of demand levels, as compared between the various GSPs (i.e. intra-group variability is lower for Mondays and Fridays).

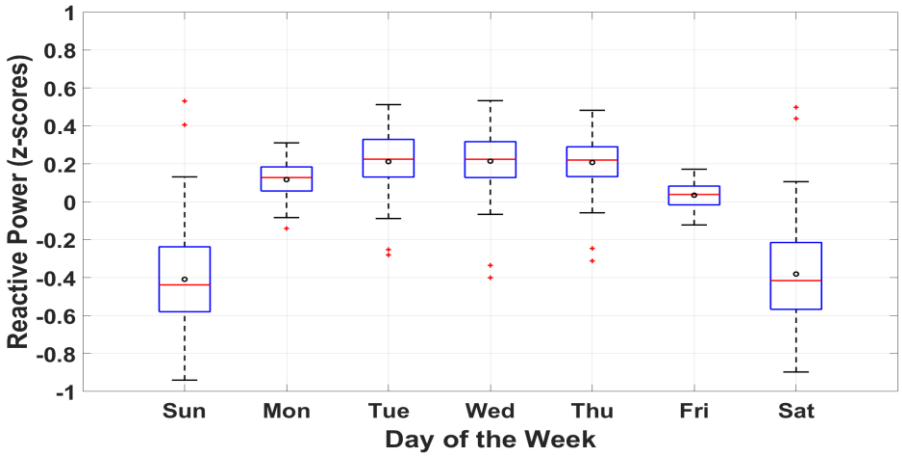


Figure 3.17: Groups of normalised reactive power for each day of the week based on 77 GSPs

The results shown in Table 3.7, indicate that Sundays and Saturdays have different reactive power levels with all weekdays and same mean and median levels with each other. The Null hypothesis is also not rejected for Tuesdays with Wednesdays, Tuesdays with Thursdays and Wednesdays with Thursdays. Note that unlike active power, there are more pairs of weekdays for which the Null Hypothesis is only marginally rejected (or not-rejected), e.g. Mondays with Wednesdays and Mondays with Thursdays. The overall agreement between the ANOVA and K-W tests is shown to be high for reactive power as well.

Table 3.7: One-way analysis of variance (ANOVA) and Kruskal-Wallis tests results, for group differences between reactive power demand for the 7 days of the week

Pair:	p-value:	Pair:	p-value:	Pair:	p-value:
Sun – Mon	3.71×10^{-8} (3.71×10^{-8})	*Mon – Wed*	0.02 (0.02)	*Tue – Sat*	3.71×10^{-8} (3.71×10^{-8})
Sun – Tue	3.71×10^{-8} (3.71×10^{-8})	*Mon – Thu*	0.04 (0.04)	*Wed – Thu*	1 (1)
Sun – Wed	3.71×10^{-8} (3.71×10^{-8})	*Mon – Fri*	0.09 (0.04)	*Wed – Fri*	7.33×10^{-8} (4.76×10^{-8})
Sun – Thu	3.71×10^{-8} (3.71×10^{-8})	*Mon – Sat*	3.71×10^{-8} (3.71×10^{-8})	*Wed – Sat*	3.71×10^{-8} (3.71×10^{-8})
Sun – Fri	3.71×10^{-8} (4.11×10^{-7})	*Tue - Wed*	1 (1)	*Thu – Fri*	1.92×10^{-7} (5.83×10^{-8})
Sun - Sat	0.97 (99)	*Tue – Thu*	1 (1)	*Thu – Sat*	3.71×10^{-8} (3.71×10^{-8})
Mon – Tue	0.02 (0.03)	*Tue – Fri*	9.84×10^{-8} (5.82×10^{-8})	*Fri - Sat*	3.71×10^{-8} (1.59×10^{-6})

Figure 3.18 shows the corresponding results for voltage. In contrast with the results for active and reactive power, weekends are characterised by higher voltage levels than weekdays, despite the fact that this variability is limited (as discussed in Section 3.3).

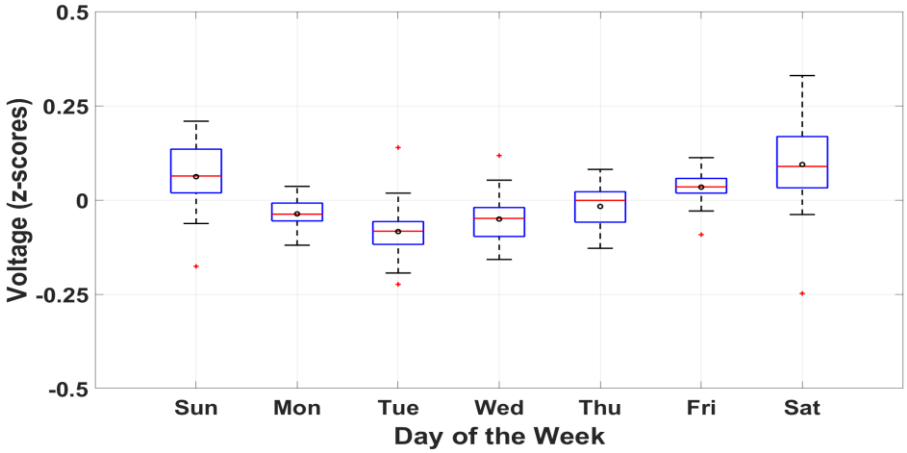


Figure 3.18: Groups of normalised voltage for each day of the week based on 24 GSPs

Higher voltage levels during weekends are (assumed) to be due to the decrease in both active and, particularly, reactive power demands. Furthermore, as most of the electrical devices and equipment in the residential-sector are with high power factors (generally greater than 0.95, which is ensured by equipment manufacturers), it is also assumed that the voltage increase during weekends is primarily due to the decreasing demands for reactive loads, used in the commercial and industrial sectors.

The Null hypothesis results for voltage, are presented in Table 3.8 and confirm the observations made about the weekday/weekend distinctions in Figure 3.18. Fridays are also shown to "cluster" with weekends, regarding the mean and median voltage levels, whereas the

Null hypothesis can be rejected between Fridays with all weekdays apart from Thursdays. The consistency of the resulting p-values between the ANOVA and K-W models, for all three parameters *P*, *Q* and *V*, indicates that the ANOVA assumptions are not violated, or at least, not violated to the extent that false-positive and false-negative errors are present.

Table 3.8: One-way analysis of variance (ANOVA) and Kruskal-Wallis tests results, for group differences between voltage levels for the 7 days of the week

Pair:	p-value:	Pair:	p-value:	Pair:	p-value:
Sun – Mon	7.25×10^{-5} (7.07×10^{-4})	*Mon – Wed*	0.99 (0.99)	*Tue – Sat*	3.71×10^{-8} (3.80×10^{-8})
Sun – Tue	3.72×10^{-8} (3.71×10^{-8})	*Mon – Thu*	0.96 (0.93)	*Wed – Thu*	0.69 (0.69)
Sun – Wed	2.69×10^{-6} (3.71×10^{-8})	*Mon – Fri*	0.02 (0.007)	*Wed – Fri*	0.001 (0.001)
Sun – Thu	0.004 (0.03)	*Mon – Sat*	5.30×10^{-8} (3.71×10^{-8})	*Wed – Sat*	3.73×10^{-8} (1.53×10^{-6})
Sun – Fri	0.85 (0.99)	*Tue - Wed*	0.71 (0.89)	*Thu – Fri*	0.19 (0.17)
Sun - Sat	0.74 (0.98)	*Tue – Thu*	0.03 (0.08)	*Thu – Sat*	3.87×10^{-6} (0.002)
Mon – Tue	0.29 (0.62)	*Tue – Fri*	6.91×10^{-7} (3.58×10^{-6})	*Fri - Sat*	0.07 (0.08)

Although the general trends of demand levels within weekly cycles have been discussed, it should be noted that the results also indicate that not all GSPs have the same "behaviour" when considering the different levels for the seven days of the week and particularly regarding the weekday/weekend distinction. This is demonstrated by the fact that the extent of intra-group variability is higher for Sundays and Saturdays compared with the weekdays, as shown in Figures 3.16, 3.17 and 3.18 (extent of whiskers in box-plots).

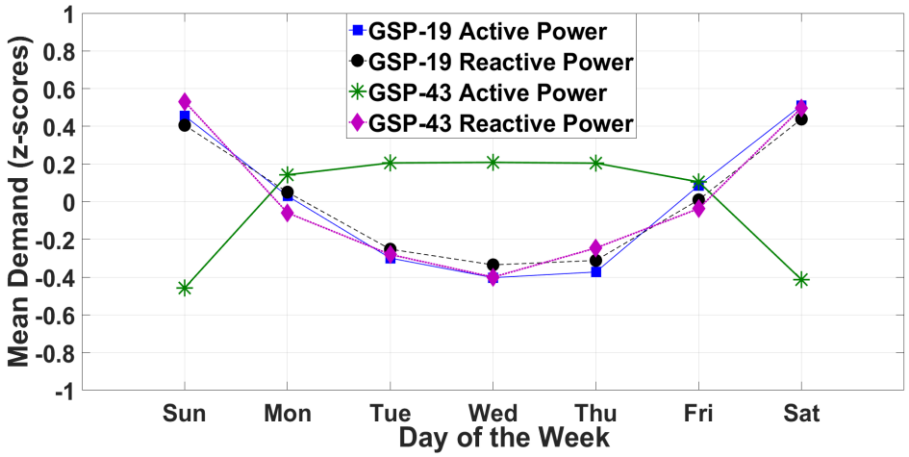


Figure 3.19: GSPs-19 and 43 which can be considered as "atypical" based on the weekly cycle analysis (as compared to the general trends)

As an example, consider Figure 3.19 which shows the mean standardised demands (*P* and *Q*) for GSPs-19 and 43. GSP-19 has higher demand levels during weekends for both active and

reactive power, while GSP-43 has higher reactive power levels during the weekends but lower active power levels, which is unusual considering the positive correlations between the two parameters. In fact, it can be shown, that the relative difference in normalised active power demand levels between weekdays and weekends, is a predictor of the percentage-contribution of domestic to total demand, for different GSPs. This is illustrated in Figure 3.20. TR (y-axis) corresponds to the total residential demand per GSP (which is constituted of ordinary and economy-7 consumptions). These percentages are estimated based on the customer-class disaggregation approach and are therefore discussed in more detail in Chapter 5.

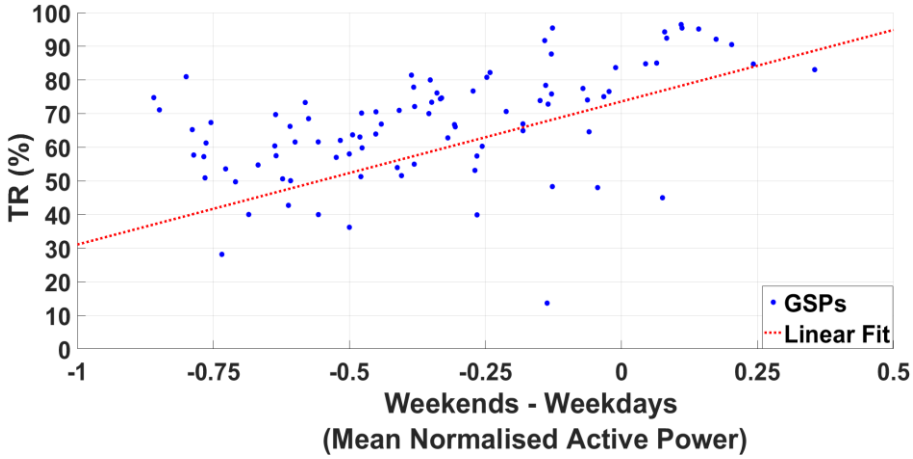


Figure 3.20: Active power demand difference (normalised values) between weekdays and weekends and percentage of total residential demand – TR (%)

Figure 3.20 shows that GSPs with higher (estimated) residential demand, as a percentage to total measured demand, have increased active power levels during the weekends, as expressed by the difference in normalised values (x-axis). Furthermore, GSPs with positive difference (i.e. higher demands during weekends), correspond to 11 out of 20 Danish GSPs and the Slovenian GSP (which is an aggregation of LV residential demands, as discussed in Section 3.1) and these are the GSPs with the highest percentages of total residential contributions. A possible explanation is that higher penetration of residential loads (at particular GSPs), causes demand levels to increase during the weekends because people spend more hours at home than during working-days. These trends have also been reported, based on the analysis of individual household consumptions, as in [106]. For primarily commercial, industrial and mixture GSPs, weekends are characterised by decreasing industrial and commercial loads and this decrease is disproportionally larger than the smaller weekend increase from residential customers, thus a total decrease in demands is observed during weekends.

The results presented in this section are taken into consideration for various purposes throughout this thesis. In particular, analysis of correlations and dependencies and customer-

class and load disaggregation approaches (Chapters 4 to 6) make distinctions between weekdays and weekends. In Chapter 7, and on order to improve forecasting performance, demand models are formulated for the seven days of the week separately and a discussion is provided for the corresponding distribution of errors. The results are also further investigated in the next section, with respect to the diurnal profiles.

3.5 Diurnal Profiling

For the analysis of diurnal (daily) profiles, statistical parameters are estimated on a per half-hour of the day basis, or for the appropriate time-scale based on the resolution of available measurements. Figure 3.21 shows basic metrics, i.e. range of variations, 25th and 75th percentiles, mean and median values, for GSP-1. These statistics essentially represent central tendencies and dispersions in demand values over the course of one calendar year, as calculated form the data measured at each half-hour of the day and are therefore indicative of the diurnal as well as the seasonal demand variabilities.

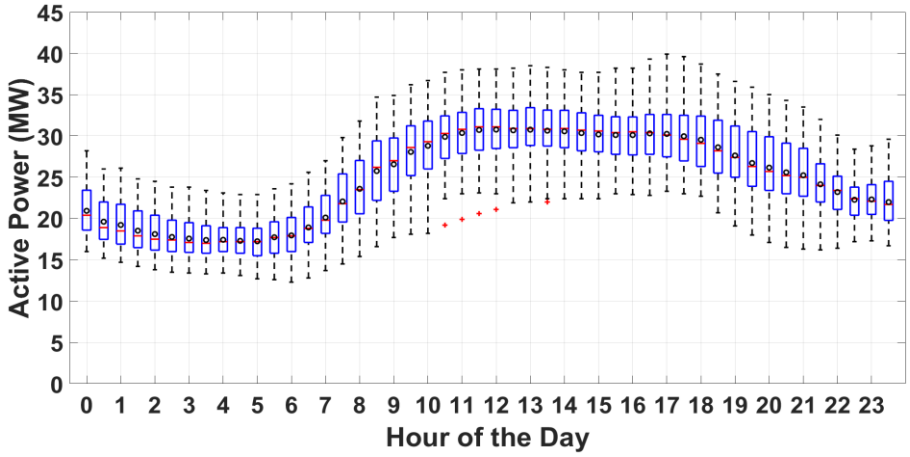


Figure 3.21: Basic statistical parameters for active power in the diurnal perspective for GSP-1

These parameters can be combined together and with various normalisation techniques in order to highlight certain features of measured demands and thus reveal patterns that can be useful for further analysis. As an example, Figure 3.22 shows four diurnal plots of active power from all 98 GSPs, each plot enhancing different characteristics of active power demand variability. Figure 3.22 (a) shows the mean values of normalised (over maximum - (3.3)) active power, which gives the characteristic mean diurnal patterns for each GSP. Figure 3.22 (b) shows the range of variations (maximum – minimum, per half hour), which is the total demand available for seasonally affected loads (such as heating/cooling and lighting loads), while Figure 3.22 (c) shows the same range normalised over the maximum demand value for the year, at each half-hour of the day (i.e. as described in Section 3.2, not global maximum, but

rather per half-hour maximum), which makes the range of variations comparable among the various GSPs but also "highlights" the variability at periods of relatively low peak demand (such as during night hours), which is useful for the classification of GSPs according to demand-percentages from economy-7 customers.

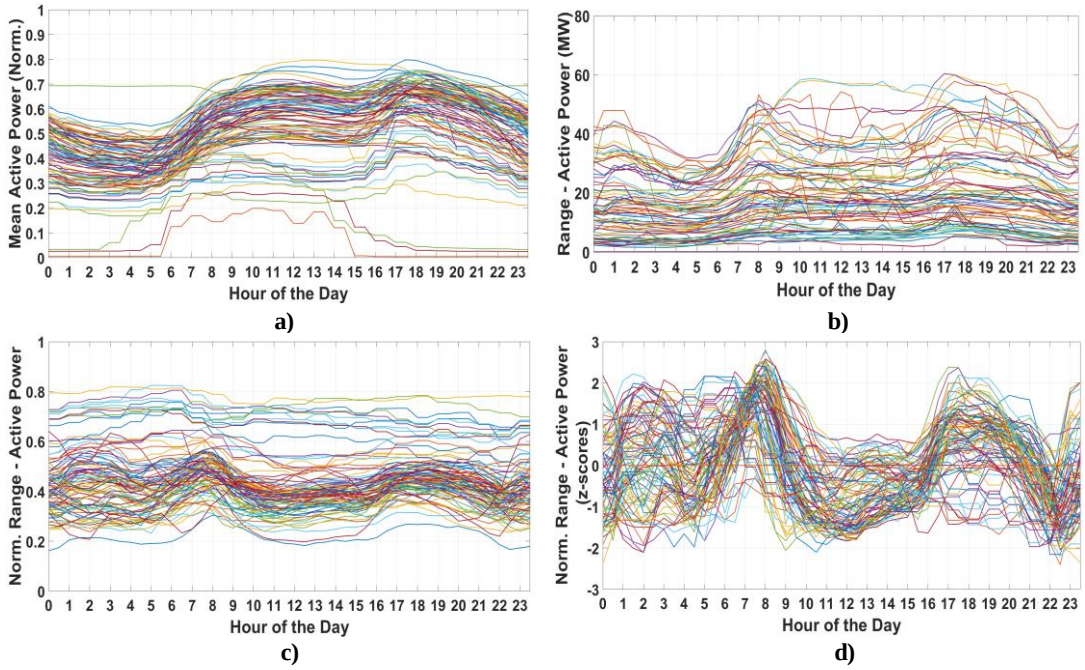


Figure 3.22: Active power for 98 GSPs: a) normalised mean demand, b) range of variations, c) normalised range of variations, d) normalised (and z-score) range of variations

Figure 3.22 (d) shows the same normalised range as in Figure 3.22 (c), but further normalised as z-scores (3.1) over the diurnal periods, to determine the relative range of variations at each half-hour with respect to the rest of the half-hours. This further highlights periods of high variability but also shows more clearly the diurnal "turning-points" for seasonal variations, e.g. around 08:00 hours at which point demand varies the most through the year (with respect to maximum demand at that period), followed by a period of lower seasonal variability before increasing again from afternoon to evening hours, between 16:00 and 22:00. These periods of "shifting" demand levels are further discussed in Section 3.8, in terms of the rate-of-change of demand, as well as in Chapter 4 in the context of seasonal sensitivities to weather conditions. The various normalisation approaches are also used as "metrics", for identifying patterns in demands, useful for customer classification in Chapter 5, as well as in Chapter 6, for determining contributions from different load-categories.

The periods of the day exhibiting similar demand levels can be determined using the non-parametric Kruskal-Wallis one-way analysis of variance test, as presented in Section 3.4. The analysis is conducted on weekdays only, based on the distinctions provided in the previous

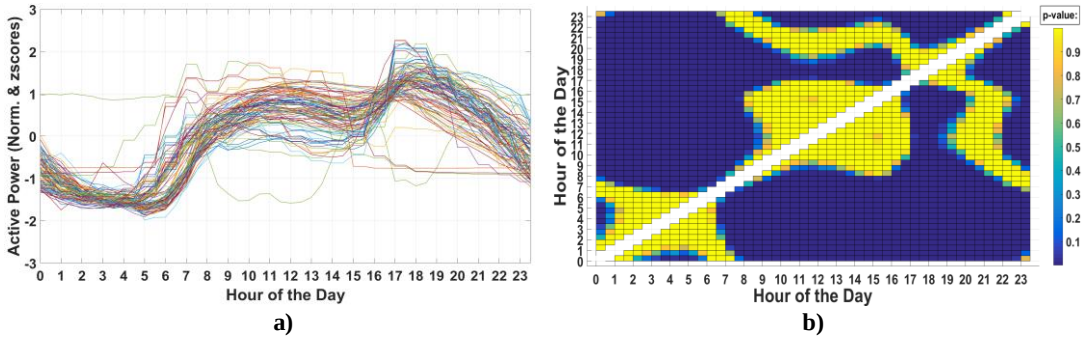
section. Differences in diurnal demand patterns between weekdays and weekends are discussed later in this section. The data preparation procedure is as follows:

- Mean diurnal profiles (weekdays) are estimated for each GSP. These correspond to the mean values as shown in the example for GSP-1, in Figure 3.21 (i.e. circles in boxes).
- For each GSP, the mean diurnal profiles are normalised according to their z-scores (3.1), to create comparable values for GSPs of different absolute demand levels.
- A 2-D array of half-hours and all GSPs is created, i.e. 48×98 for active power, 48×77 for reactive power and 48×24 for voltage.

The Null hypothesis is formulated as in (3.9), but now groups correspond to the 48 half-hours of the day, i.e. $j = 1,2,\dots,48$ and lower p-values indicate that there are substantial group differences between demand levels for the corresponding pairs of half-hours.

Figures 3.23 (a), 3.24 (a) and 3.25 (a) show the diurnal profiles for active power, reactive power and voltage, normalised using the approach discussed above. Figures 3.23 (b), 3.24 (b) and 3.25 (b) show the results for the Kruskal-Wallis tests. Due to the large number of resulting pairs, i.e. $\binom{n}{k} = 1128$, the p-values are represented graphically, instead of using tables as in Section 3.4. These are illustrated using a colour-scale, where dark blue represents p-values closer to zero (Null hypothesis can be rejected) and bright yellow for p-values closer to 1 (Null hypothesis cannot be rejected).

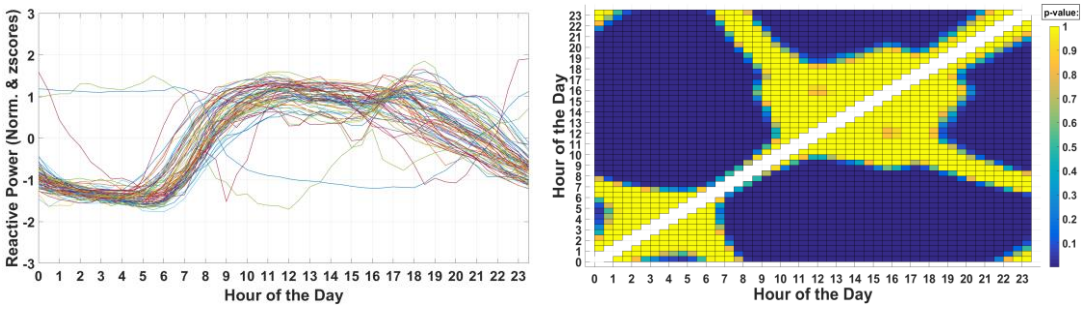
The results for active power indicate that there are similar normalised demand levels during the night, i.e. from 01:00 to 07:00 hours and for periods between 09:00 and 17:00 hours. For the period between 19:00 and 23:30 hours, the demand levels are not similar with the adjacent half-hours but rather with demand levels ranging from 07:00 to 17:00 hours (e.g. mean demand at 21:00 is similar with mean demand at 10:00).



a) **b)**

Figure 3.23: a) Normalised diurnal profiles and b) Null hypothesis test results for group differences in mean demand levels among the 48 half-hours of the day, using values from 98 GSPs, for active power

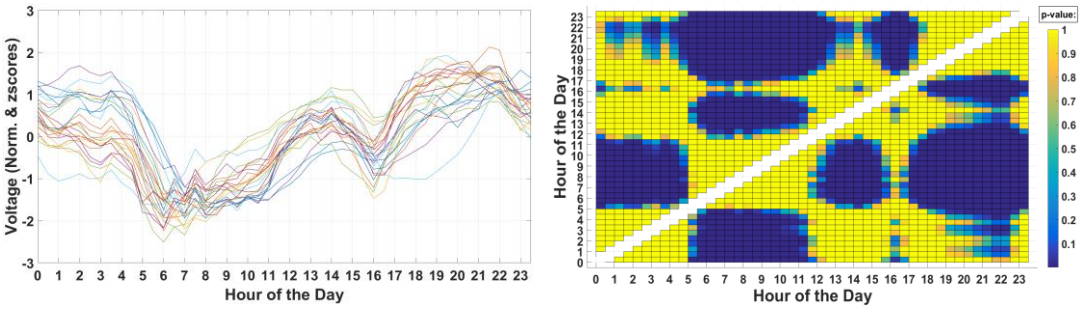
For reactive power, Figure 3.24, normalised demands are at approximately the same level for night periods between 01:00 to 07:00 hours (as in the case of active power) and between 10:00 and 20:00 hours (this period extends further than in the case of active power). Similarities also exist between the evening hours from 18:00 to 23:30 and the morning hours from 06:00 to 10:00.



a) **b)**

Figure 3.24: a) Normalised diurnal profiles and b) Null hypothesis test results for group differences in mean demand levels among the 48 half-hours of the day, using values from 77 GSPs, for reactive power

Voltages are shown in Figure 3.25 and are generally at similar (normalised) levels for periods of the day between 18:00 and 04:00 hours, between 05:00 to 12:00 hours and between 12:00 to 18:00 hours. There are also similarities between non-adjacent periods such as between night hours and afternoon hours. The results in Figure 3.25 (a) also show that voltage levels reach minimum values during early morning hours (between 05:00-08:00) with a second trough at around 16:00 hours and maximum levels during evening and night periods.



a) **b)**

Figure 3.25: a) Normalised diurnal profiles and b) Null hypothesis test results for group differences in mean voltage levels among the 48 half-hours of the day, using values from 24 GSPs

The presented analysis takes into account demands from all available GSPs, which are, however, expected to vary for single GSPs or for specific groups of GSPs (such as GSPs supplying commercial and/or industrial demands). Normalised diurnal profiles with respect to customer-sectors are presented and discussed in Chapter 5.

The group differences between demand levels for weekdays and weekends (Section 3.4) can be further investigated using the diurnal profiles, as shown in an example in Figure 3.26 (a) and (b), for active power and reactive power respectively, for GSP-20. Normalisation for this analysis is performed with respect to the maximum value per dataset (3.3).

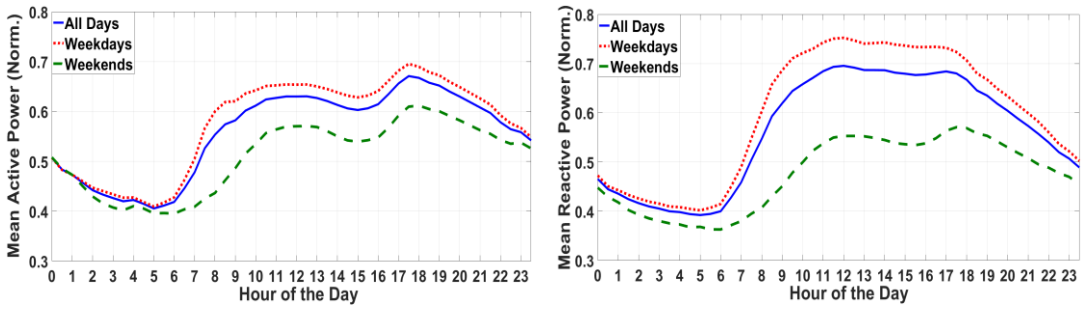


Figure 3.26: Weekday/weekend mean normalised demand levels for: a) active and b) reactive power

The differences can be quantified for all GSP, by subtracting the mean-normalised demand for weekends from the mean-normalised demand for weekdays, as shown in Figures 3.27 (a) and (b) for active power and reactive power. This approach allows to determine the exact periods of the day, responsible for the weekday/weekend group differences discussed in Section 3.4, as well as to quantify their extent in terms of normalised values.

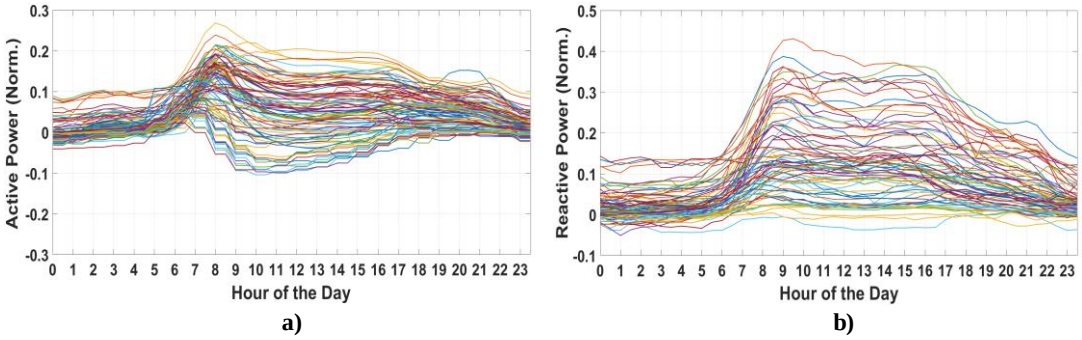


Figure 3.27: Normalised demand differences between weekdays and weekends for all available GSPs: a) active power and b) reactive power

The higher levels of normalised differences for reactive power compared with active power (between weekdays and weekends), as show in Figure 3.27, can be explained by the stronger weekly components of reactive power as demonstrated in Section 3.3, Table 3.5. The current results show the diurnal periods for which this weekly component is stronger, i.e. between 08:00 to 17:00 hours, for the majority of GSPs. For active power, higher differences are found during morning hours and particularly between 07:00 to 10:00, which is also reflected in the particular normalisation approach presented in Figure 3.22 (d). There are also GSPs with higher levels of active power demand on the weekends (as mentioned in Section 3.4, for more than half of the Danish GSPs) and the differences are concentrated between 08:00 and 17:00

hours. This further supports the hypothesis that these GSPs are supplying predominantly residential customers and that the reason for increased demands during weekends, at these particular diurnal periods (i.e. morning to afternoon hours), is that customers are at work during weekdays.

Therefore, the differences in demand levels, as presented from the diurnal perspective, with distinctions made between weekdays and weekends, can be useful in determining daily periods that are primarily affected by loads related to behavioral patterns (people's schedules) and occupancy levels, as opposed to seasonally affected (e.g. weather related) loads. Consider for example, the normalised demand differences (active and reactive) between weekdays and weekends, for a predominantly residential and a commercial/mixture GSP (i.e. GSP-14 and GSP-3, respectively), shown in Figure 3.28.

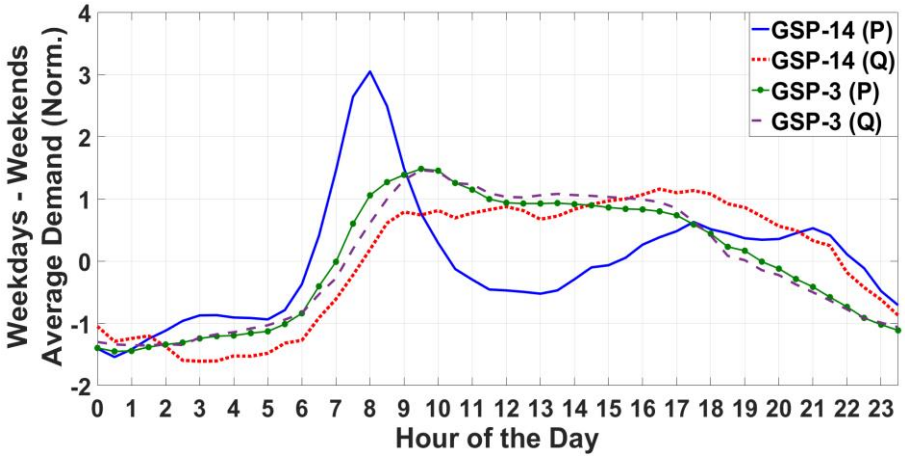


Figure 3.28: Active and reactive demand differences between weekdays and weekends for GSPs-14 & 3

The morning peak in the weekday-weekend demand difference, comes earlier for the residential GSP than for the commercial GSP and it is more pronounced. Furthermore, the diurnal patterns of *P* and *Q*, for the commercial GSP, correlate almost perfectly, whereas for the residential GSP they do not and the morning and evening peaks in active power do not coincide with peaks of reactive power. This indicates that the portion of the load characterised by the morning peak (and to a lesser extent for the evening peak), corresponds to higher contributions from mostly-resistive loads and can include water-showers, cooking resistive-elements, heating (programmable thermostats), etc. Patterns of use for such appliances are reported in literature, as in [106] and are also used for constructing component-based load models, as in [188]; and are in agreement with the profiles shown in Figure 3.28.

Analysis on a per half-hour (or other available resolution) basis also enables profiling of seasonal components for specific diurnal periods and not as average values for each day of the

year and these distinctions can be used when considering the inputs for correlation and regression analysis. Furthermore, instead of determining group differences between individual half-hours (as in this section), the diurnal profiles can be used to make distinctions in consumption patterns among various GSPs and thus allow for the classification of primary/secondary substations. Diurnal profiles can also be used to make distinctions between peak, intermediate and base demands and have applications in the disaggregation analysis. All the above are therefore discussed in more detail in the subsequent chapters.

3.6 Seasonal Profiling

Figure 3.29 shows the basic statistical parameters for active power, reactive power and voltage, for each month of the year and for weekdays only, based on data from all available GSPs (i.e. 98 for *P*, 77 for *Q* and 24 for *V*).

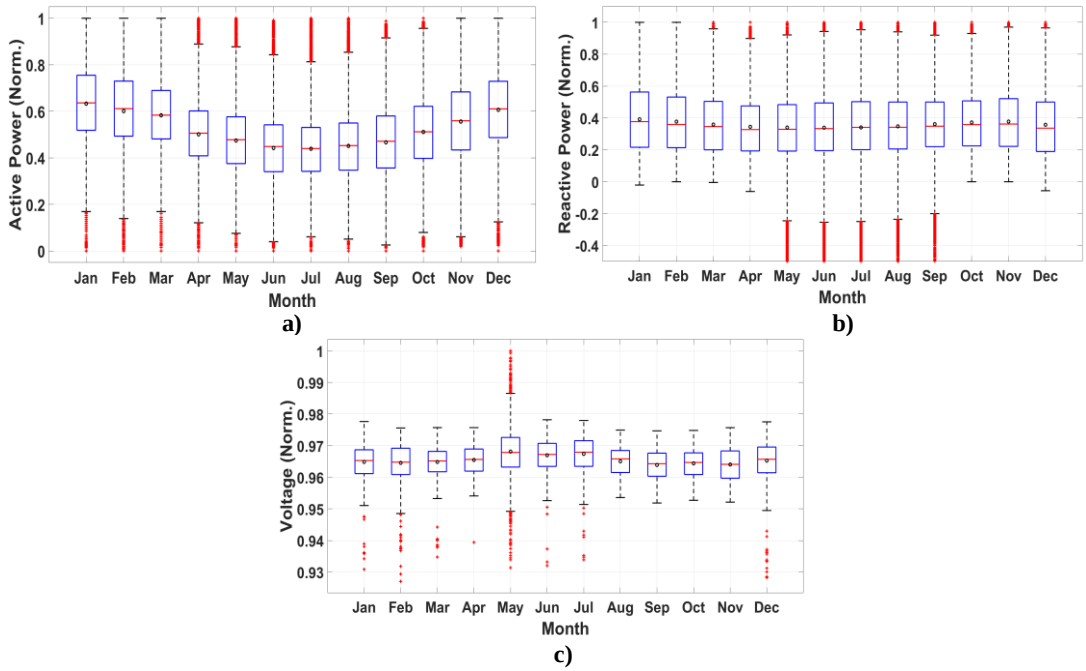


Figure 3.29: Basic statistical parameters, for all available GSPs for: a) active power, b) reactive power and c) voltage, in the seasonal perspective

The seasonal component is mostly evident for active power, which, as presented in Section 3.3, has the highest normalised magnitude as well as the highest percentage-contribution to the total range of variations, for seasonal (yearly) cycles, compared to reactive power and voltage. However, for individual GSPs or groups of GSPs with common characteristics (such as with primarily residential consumption), seasonality would also be shown for reactive power, but in Figure 3.29 (b) these are "averaged-out" by the differences in the seasonal periodicities from the various GSPs. This diversity in the extent of the reactive power seasonal

component is also shown in Table 3.5, in terms of the standard deviation of the percentage to total variability, which is at levels comparable to the mean.

In Figures 3.30 (a), 3.31 (a) and 3.32 (a), the seasonal profiles of active power, reactive power and voltage are presented, for all available GSPs. Because seasonality includes the hourly, daily and weekly variations, it is necessary to apply data-filtering in order to reduce the effect of these components and thus more clearly represent the yearly cycle. The "smoothed" values are calculated using a moving-average filter of approximately ± 2 weeks, i.e. 20 weekdays, in the form of:

$$X_{MA_i} = \frac{\sum_{i-n}^{i+n} X_i}{n+1} \qquad (3.13)$$

where X_i is the mean normalised (3.3) demand at weekday i and X_{MA_i} is the moving-average value at weekday i for a window length of $n = 20$, including the center value of i. Values at the start/end-points of each array are connected ("looped"), as to not compromise the filtering procedure due to positive/negative biases. The resulting seasonal curves are further normalised with respect to the 261 weekday values (using (3.1)), to produce comparable individual curves among GSPs. In Figures 3.30 (b), 3.31 (b) and 3.32 (b), the results of the K-W test (Section 3.4) regarding group differences between pairs of weekdays of the year are presented and using the same colour-scale as for the diurnal analysis of Section 3.5.

For active power, Figure 3.30 (a), there is a high percentage of GSPs clustering around the "expected" seasonality pattern, with decreased demand levels during the summer months and demand levels higher during the winter months (up to 2 standard deviations from the mean in both cases).

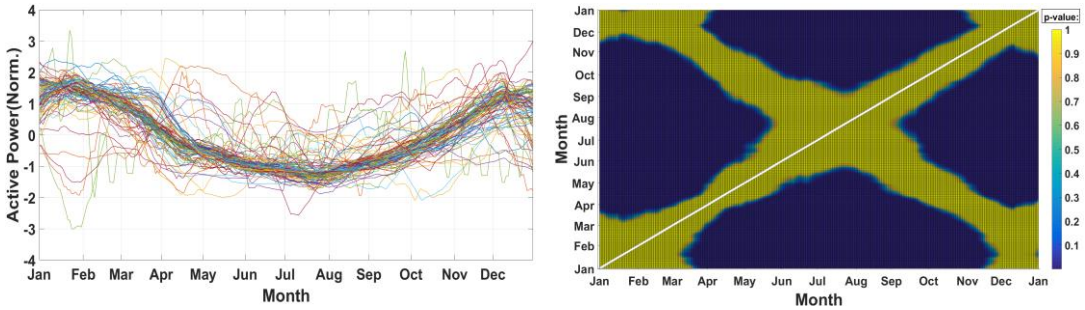


Figure 3.30: a) ''Smoothed'' and normalised seasonal profiles and b) Null hypothesis test results for group differences in mean demand levels among the weekdays of the year, using values from 98 GSPs, for active power

Similar demand levels, Figure 3.30 (b), can be expected for the summer months as well as for the yearly periods on both sides of the minimum demand levels, demonstrated by the symmetrical distribution of demands, which have a central through point at around mid-July

to August (i.e. similar demand levels for May and October, for April and November, etc.). This indicates an almost exact 12-month periodicity in variations of active power demands. A discussion about the phase of this periodicity compared to the corresponding periodicities in weather parameters and the resulting effects on electricity demands, is given in Chapter 4.

The same "clear" clustering of demands around a specific seasonality pattern is not evident for reactive power, in Figure 3.31 (a), with a number of GSPs showing no distinctive periodicities and with some GSPs actually having higher reactive power demand levels during the summer months. This is also reflected in the results for the group differences which are shown in Figure 3.31 (b), as well as in the Fourier analysis results, as discussed in Section 3.3.

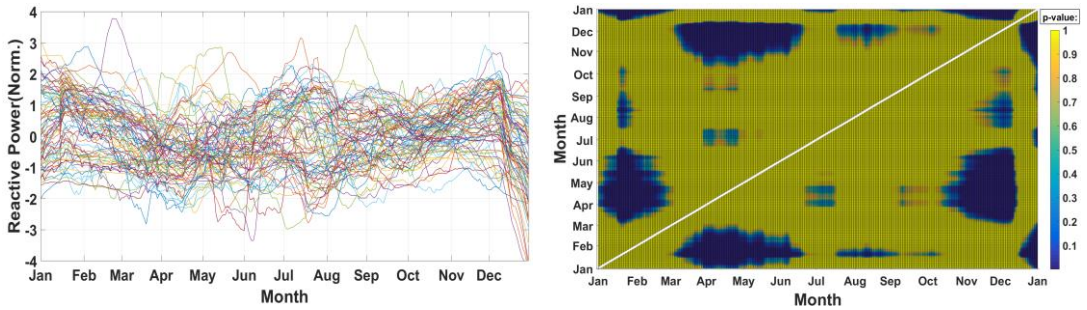


Figure 3.31: a) ''Smoothed'' and normalised seasonal profiles and b) Null hypothesis test results for group differences in mean demand levels among the weekdays of the year, using values from 77 GSPs, for reactive power

For voltage, Figure 3.32 (a), about half of the available GSPs display decreased levels during the summer period with the rest showing more erratic patterns with no distinctive periodicities. Figure 3.32 (b) indicates that the Null hypothesis cannot be rejected for, almost, all pairs of weekdays of the year (i.e. p-values close to 1, for most periods, as in the case of reactive power demands), primarily due to the effects of the GSPs that display no distinctive seasonality.

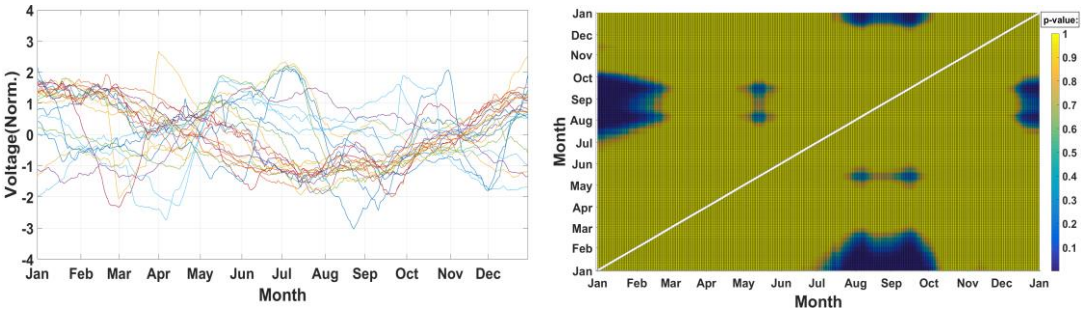


Figure 3.32: a) 'Smoothed' and normalised seasonal profiles and b) Null hypothesis test results for group differences in mean demand levels among the weekdays of the year, using values from 24 GSPs, for voltage

An analysis of the GSPs that do and do not show voltage seasonality with respect to the disaggregated percentages from customer-classes (Chapter 5), has revealed no association

between the two distinctions, i.e. the GSPs with voltage seasonality do not correspond to GSPs with higher or lower contributions from total residential or industrial and commercial demands. However most of the GSPs with available voltage measurements, have demands from mixtures of sectors, i.e. between ~50 % to ~70 % total residential demand, with very few exceptions. Further analysis is necessary to determine the reasons for the voltage seasonality, which is, as mentioned before, generally below ± 6 % of the nominal.

Representations of an individual GSP are shown in Figure 3.33, using the mean, maximum and minimum daily values (weekdays only), as well as the corresponding moving-average values (3.13) for GSP-14, for active (a) and reactive (b) power demands.

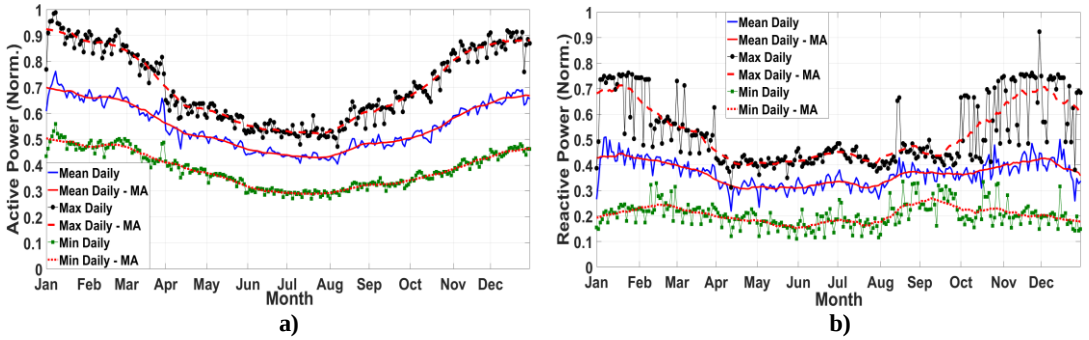


Figure 3.33: Seasonal variations: daily mean, maximum and minimum values (weekdays only) for a) active power, and b) reactive power, for GSP-14

These profiles can be considered as the "signature" seasonality profiles of each MV-GSP and are used for an initial seasonal correlation analysis (Chapter 4), which is then expanded into a seasonal per half-hour of the day analysis. An example of the seasonal variations, at specific diurnal periods, is shown in Figure 3.34 for normalised and normalised moving-average (3.12) active power values, for four selected hours of the day (i.e. 03:00, 09:00, 15:00 and 21:00), for GSP-47 and considering weekdays only.

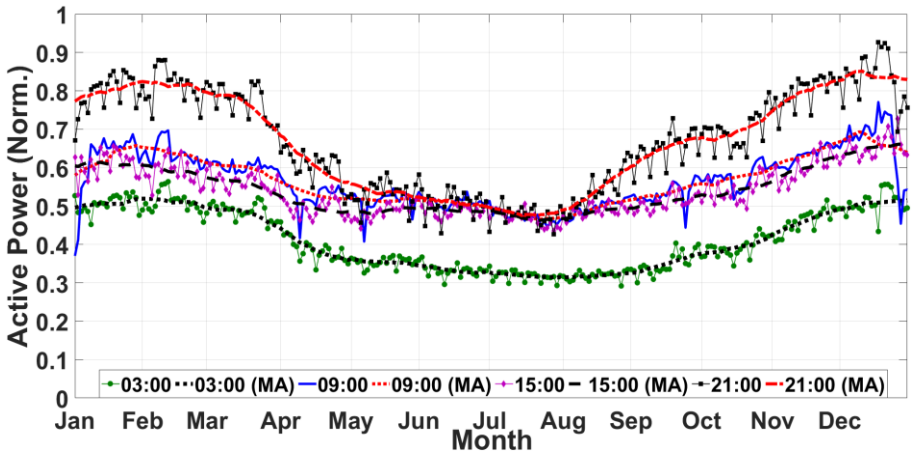


Figure 3.34: Normalised and normalised moving-average values for active power at selected hours of the day, for GSP-47 (weekdays only)

By combining the diurnal and seasonal perspectives, a "full-scale" demand profile is presented in Figure 3.35, for active power demand at GSP-53, for all days of the year.

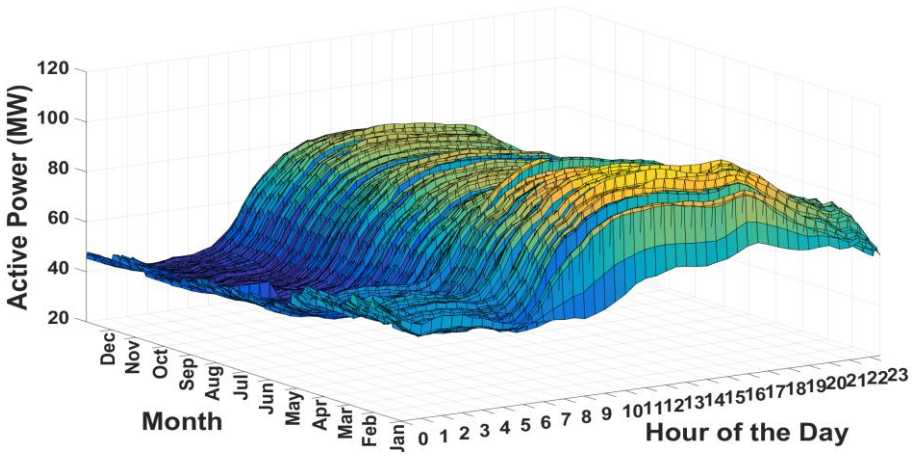


Figure 3.35: Combined diurnal and seasonal profiles for active power demand, GSP-53, all days

As in the case of diurnal profiles, the combinations of different normalisation, averaging and filtering approaches produce seasonal profiles that can be used for further studies, e.g. as inputs for regression analysis and for clustering, classification and disaggregation purposes. Accordingly, the demand representations discussed in this section are used in various instances in the following chapters of this thesis.

3.7 Occurrence of Maximum and Minimum Demands

Figures 3.36, 3.38 and 3.40 present the probability of occurrence of maximum (peak) and minimum demands per half-hour of the day for active power, reactive power and voltage, respectively. Figures 3.37, 3.39 and 3.41 show the corresponding occurrences of maximum and minimum values, with respect to the diurnal as well as the seasonal time-frames. Distinctions are, again, made between weekdays and weekends due to the different demand levels between the two groups, as discussed in Sections 3.4 and 3.5.

The analysis for active power (Figure 3.36), shows no distinctive differences in the time of occurrence of maximum and minimum demands between weekdays and weekends. Therefore, even though the demand levels between these two groups are different (for both mean and median values, as shown in Section 3.4, but this also applies for maximum and minimum values), the diurnal periods for which these maximum and minimum demands occur are shown to almost, but not fully coincide. Maximum demand occurs, primarily, between 16:00 to 22:00 hours, as well as between 09:00 to 14:00 hours. Minimum demand values are clustered in one

main group between 00:00 to 06:00 hours, although this shifts to around 08:00 for weekends (people tend to wake-up later during weekends).

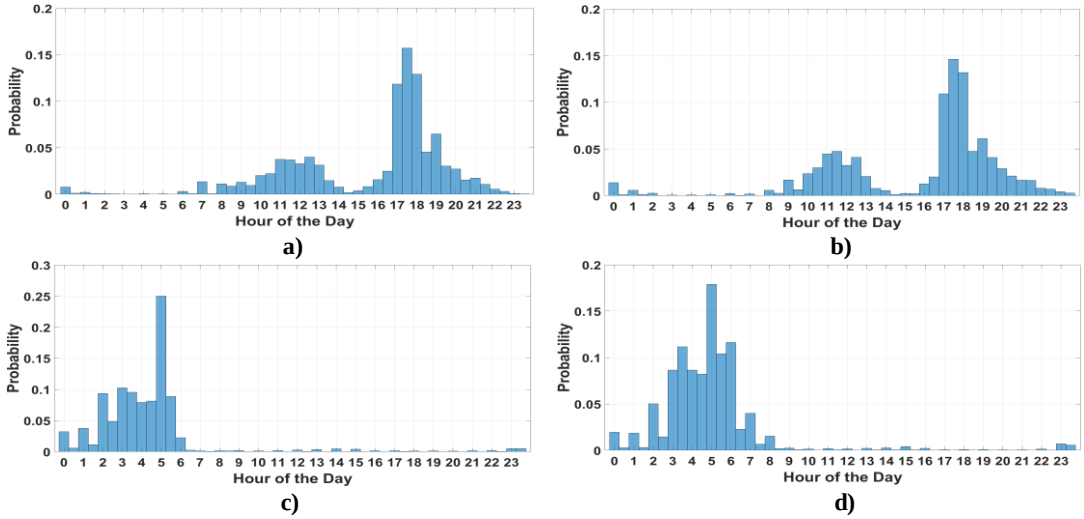


Figure 3.36: Probability of occurrence of maximum demand for: a) weekdays, b) weekends and minimum demand for c) weekdays, d) weekends, for active power

The results are based on the analysis of all available GSPs and thus no distinctions have been made between residential, commercial, industrial or mixed-type GSPs. The two peaks in the probabilities for maximum demand can therefore, at least partly, be attributed to the differences in the diurnal consumption patterns between GSPs supplying different customer-sectors. Another reason for the two clusters is the change in the time of the maximum demand occurrence for different seasons. The influence of both time-frames is illustrated in Figure 3.37, for the English GSPs and for weekdays only (the smaller dataset was selected to make the resulting patterns more distinguishable).

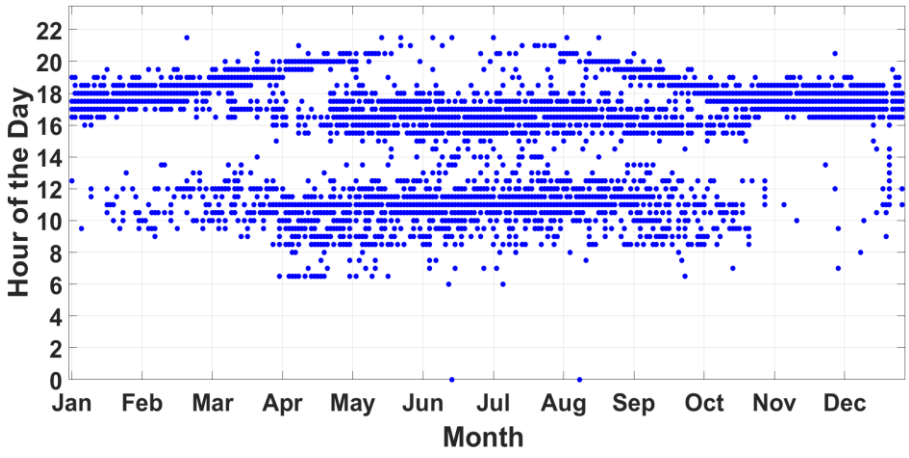


Figure 3.37: Maximum (peak) demand occurrence per half hour of the day, per month of the year (weekdays only), for active power

The data-points in Figure 3.37 mark the time of the day, for each weekday of year, at which maximum demand for active power occurs. There are two characteristic upper and lower clusters that are, primarily, due to peak demands from residential (upper) and industrial and commercial (lower) GSPs. Furthermore, there is a clear tendency of increase of the time at which maximum demand occurs, for the upper band, from winter to spring (until mid-May), followed by a drop around June, from evening hours to afternoon hours and mid-day hours until mid-August to September. Possible explanations include the shift of active power demands for heating and lighting loads to later hours of the day as the summer period approaches (increase in temperature and solar irradiance levels), up to the level where peak demand no longer occurs during evening hours because of minimum heating/lighting demand and therefore, the peak is now situated during the afternoon hours. As mentioned before, the analysis of characteristic demand patterns for GSPs supplying difference mixtures of customer-classes is presented in Chapter 5. The effects of external variables such as temperature levels and particularly, for the pattern shown in Figure 3.37, solar irradiance, are presented in Chapter 4.

Figure 3.38 shows the probability of occurrence of maximum/minimum demands for reactive power, for weekdays and weekends. Similar to the results presented for active power, there are no clear distinctions between these periods when comparing the two groups.

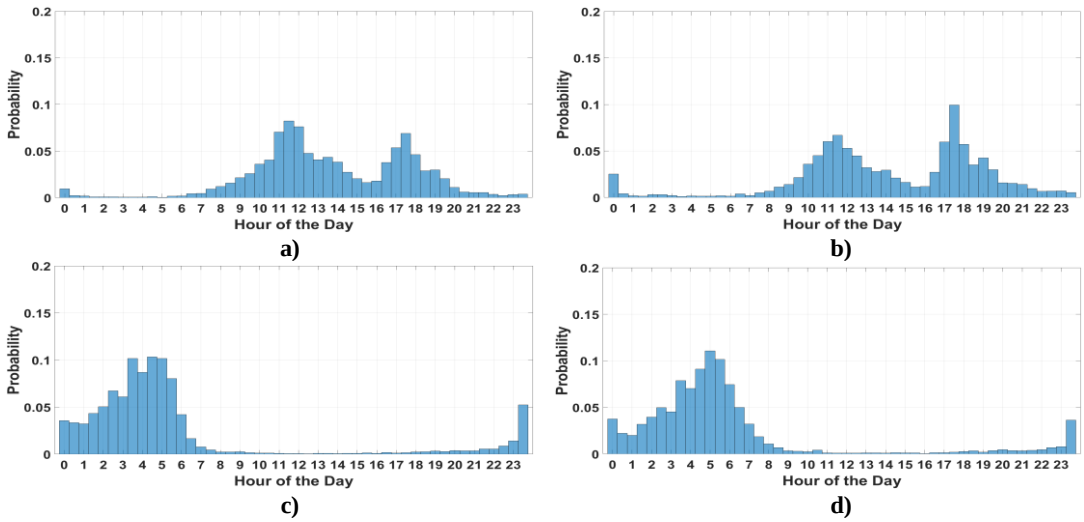


Figure 3.38: Probability of occurrence of maximum demand for: a) weekdays, b) weekends and minimum demand for c) weekdays, d) weekends, for reactive power

Maximum reactive power demand occurs between 16:00 to 22:00 hours and 08:00 to 15:00 hours, but unlike active power, peak demands for reactive power have a higher probability of occurrence during the mid-day period, comparable, or in the case of weekdays higher, than during the evening period. This shows a higher penetration of reactive load components in the

system during this period, which can be attributed to a higher percentage of commercial and industrial loads around mid-day. Minimum demand levels are, again, mostly concentrated around the night and early morning periods between 23:30 to 06:00, with a slight shift in minimum reactive power demand for weekends, up to 08:00 hours (as in the case of active power).

In Figure 3.39 the periods of occurrence of maximum reactive power demand as a function of hour of the day and weekday of the year are shown. Although the upper and lower bands are evident, as in the case of active power, the lower band (mid-day) is far more populated during the summer period (possible increase of air-conditioning loads from the commercial sector and a general decrease in reactive power demand during evening hours from the residential sector).

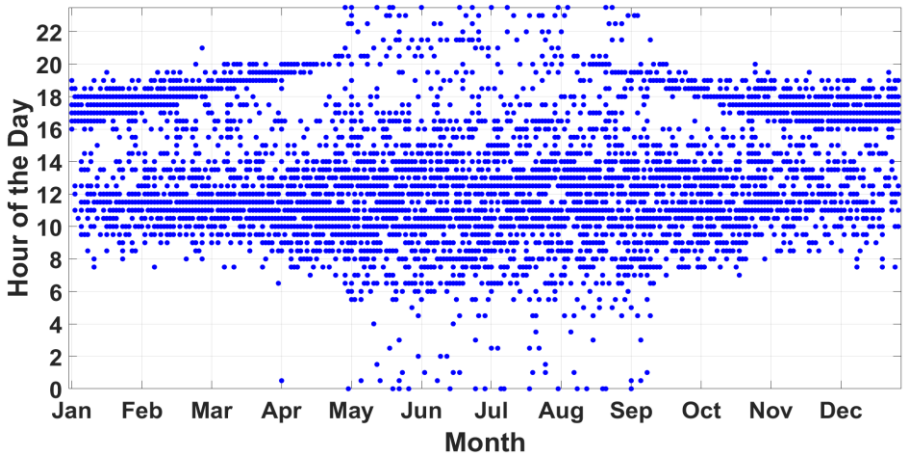


Figure 3.39: Maximum (peak) demand occurrence per half hour of the day, per month of the year (weekdays only), for reactive power

Figure 3.40 shows the probability of occurrence of maximum/minimum voltage levels. No significant differences are again shown between weekdays and weekends, apart from a slight increase in the probability of maximum voltage occurrence during mid-day for weekends. The highest probabilities for maximum voltage levels are found at the periods between 17:00 and 00:00 hours, although these extend to early morning hours (up to 05:00), and lower probabilities are shown between 11:00 to 15:00 hours (also reflected in the mean normalised diurnal profiles in Section 3.5). Minimum voltage levels occur primarily between 04:00 and 12:00 hours, with lower probability peaks between 15:00 and 17:00 hours.

Even though there is a higher dispersion of maximum voltage levels per half-hour of the day over a one calendar year, as shown in Figure 3.41, as in the case of active and reactive power demands there is a clearly identifiable pattern of shifting peak levels from early afternoon to evening hours (i.e. shift from 16:00 to 22:00 hours).

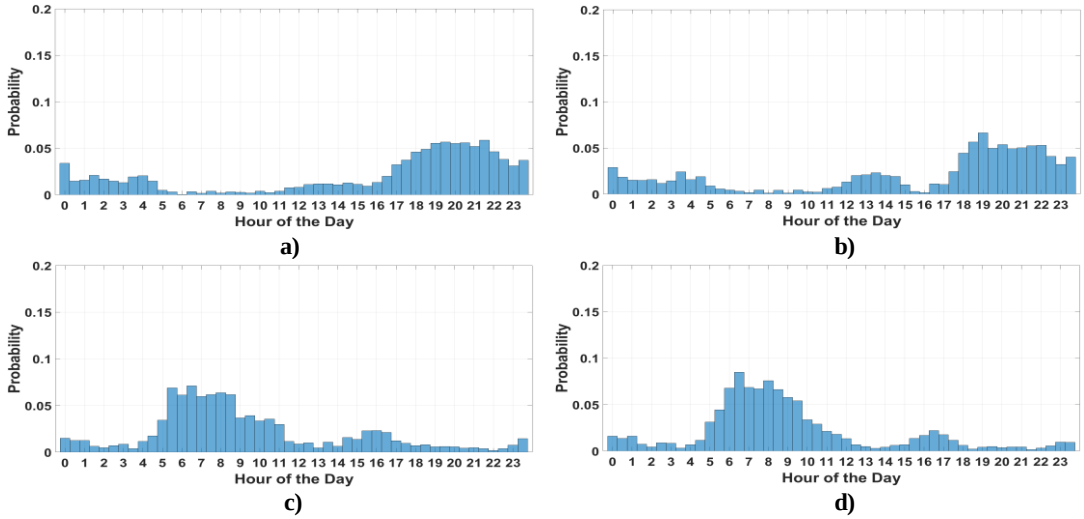


Figure 3.40: Probability of occurrence of maximum voltage for: a) weekdays, b) weekends and minimum voltage for c) weekdays, d) weekends

These patterns have strong correlations with "markers" of seasonality, such as the solar position and solar irradiance levels (further discussed in Chapter 4). For example, maximum voltage levels shift to later diurnal periods during summer months, with the maximum shift found around mid to late June, which is the period of maximum solar elevation angles (directly related to solar irradiance and temperature levels, i.e. summer solstice).

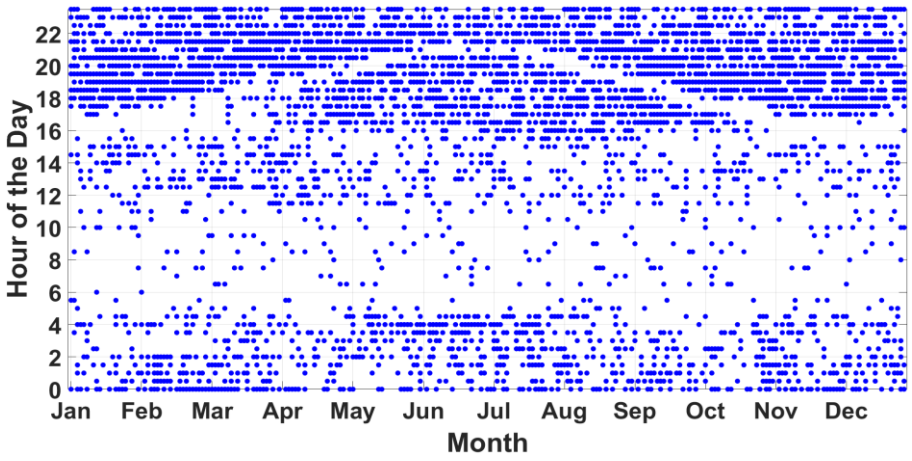


Figure 3.41: Maximum (peak) voltage occurrence per half hour of the day, per month of the year (weekdays only)

The results presented in this section are based on computational/statistical analysis that can easily be adjusted to accept inputs from single GSPs or groups of GSPs from particular locations and DNOs. Knowledge of maximum and minimum demand levels can facilitate demand side management schemes, but these approaches are often restricted by limited dimensionality with respect to temporal perspectives and are often based on conclusions drawn from customer-survey or household-metering analysis. Expanding maximum/minimum

demand profiles to diurnal and seasonal time-scales can accommodate dynamic approaches to peak demand shifts and by drawing conclusions directly from MV-datasets, these interventions could be tailored for the specific locations, customers, network characteristics and DNOs.

3.8 Profiling the Rate of Change of Demands

In this section, demand profiles are presented with respect to the numerical differences between adjacent measurements in the form of:

$$\Delta y_i = y_{i+1} - y_i \qquad (3.14)$$

where Δy_i is the difference in measured active power (or reactive power) between two consecutive values; current value, y_i, and the next value, y_{i+1}. This difference is therefore an equivalent of the numerical differentiation (or finite difference) for each measured value, within a time-step defined by the dataset's resolution.

These differences are calculated for the input data (as presented in Section 3.1), normalised over the maximum of each dataset by (3.3), with no further filtering or increase in dimensionality (i.e. arrangement into matrices of diurnal and seasonal time-frames) applied prior to the calculation. The resulting datasets are therefore representative of the increase/decrease in demands from one half-hour to the next (or as it applies to datasets of different resolutions). The results can then be arranged in a combined seasonal and diurnal perspective, to illustrate these changes over a one calendar year. Differences in adjacent demand values are shown as a percentage to the yearly peak demand, to produce estimates with respect to maximum expected loads, per GSP. However, other estimations are possible, with respect to other parameters or more restricted time-scales, such as Δy_i as a percentage of mean yearly demand, or as a percentage of mean demand estimated within a moving-average window (3.13) of ± 48 half hours, which would filter-out the seasonality in Δy_i values (no longer based on a common scale), but would pronounce the differences within the diurnal cycles. The examples shown in Figures 3.42 and 3.43 are for active and reactive power and correspond to GSP-5.

Increase in active power demand is noticeable as a consistent positive change throughout the year during morning hours, between 05:00 and 08:00, but interrupted during the weekends (i.e. in Figure 3.42, 52-distinctive "bands", separable by weekends). Lower but consistent positive changes are also shown throughout the year for the diurnal periods between 15:00 and 17:00 hours. There is a third "band" of increasing active power demands which shifts

throughout the year, starting from 16:00 hours during winter months and extending up to 22:00 hours during the summer months (with a peak in late June).

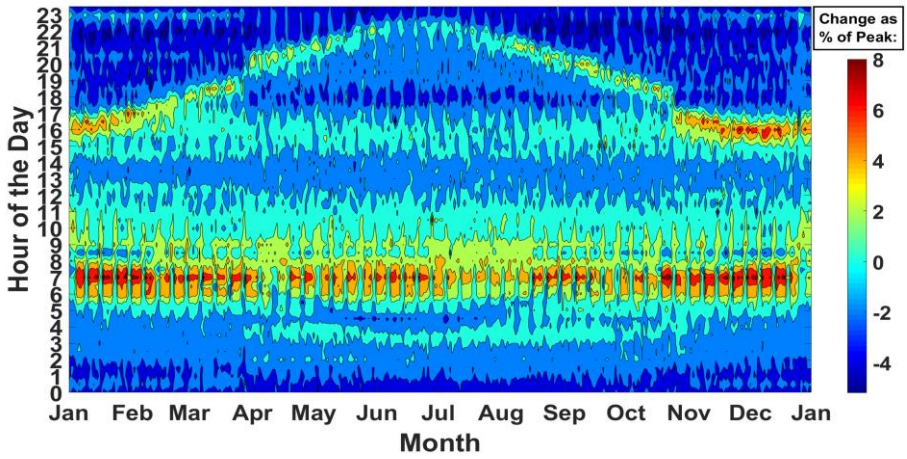


Figure 3.42: Changes in active power as a % of peak demand, GSP-5

The same components can be seen for reactive power in Figure.3.43. In particular, the upper increasing "bands" directly reflect solar irradiance levels, as the peaks coincide with the summer solstice and therefore increasing demand levels during these periods maybe directly related to lighting loads. The hypothesis is further supported by the fact that, similar upper "bands" are less evident for GSPs with primarily commercial and/or industrial consumptions. The trends may also be related to street-lighting, for which the switch-on periods are variable and controlled by photo-sensitive switches, reflecting available sunlight levels throughout the year.

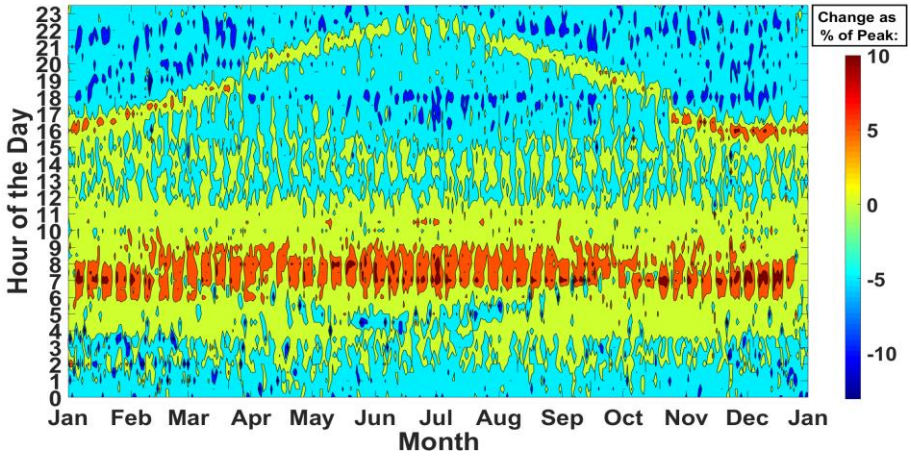


Figure 3.43: Changes in reactive power as a % of peak demand, GSP-5

In Figure 3.44, the results for active power are presented for selected months (i.e. January, March, May, July, September, November), where the afternoon-evening peaks in demand changes are highlighted. The evening peaks in active/reactive power demands (presented in

Section 3.7) result from a positive rate of change from around 15:00 to 16:00 hours (for the examples presented here), which means that during periods of maximum diurnal demand the trend for the rate of change is actually decreasing (i.e. maximum or turning point is reached and demand begins to drop). These negative changes extend through the night and up to 04:00 hours, however, for November, January and March the increasing/decreasing demands balance-out and the range of change reaches a zero value between 22:30 and 23:30 hours. This is assumed to be related to lighting loads and possibly to the time of commencement of night tariffs for heating loads. The changes with respect to external variables and the presence of heating and lighting loads are discussed in more detail in Chapters 4 and 6. The time of commencement for economy-7 customers is discussed in Chapter 5.

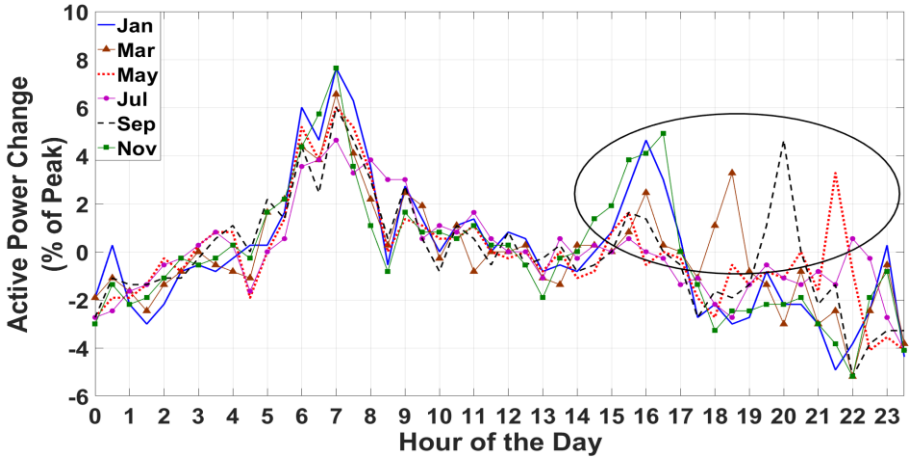


Figure 3.44: Average changes in active power demands as a percentage of peak demand for selected month, GSP-5

Profiling demands in terms of the rate of change has the potential benefit of capturing time-intervals for which the primary sources of stresses in the distribution networks cannot be accounted for by the actual system loadings and may be better assessed by the rate of change of the loading conditions. Authors in [189] and [190] present a theoretical interruption model, based on recordings of short-term and long-term interruptions, from two distribution networks in Europe, which are presented in Figure 3.45 (a). The morning peak shown for the short- and long-term interruptions, between 09:00 to 11:00 hours in (a), does not coincide with the morning peak in the rate of change of active power, between 06:00 to 08:00 hours in (b) and it actually coincides with actual measured demands at the same periods, as presented in the diurnal profiles in Section 3.5, Figure 3.23 (i.e. morning peak for active power). However, the second peak in long-interruptions (black-solid line in (a)) coincides with the, positive, peak in the rate of change of active power in (b).

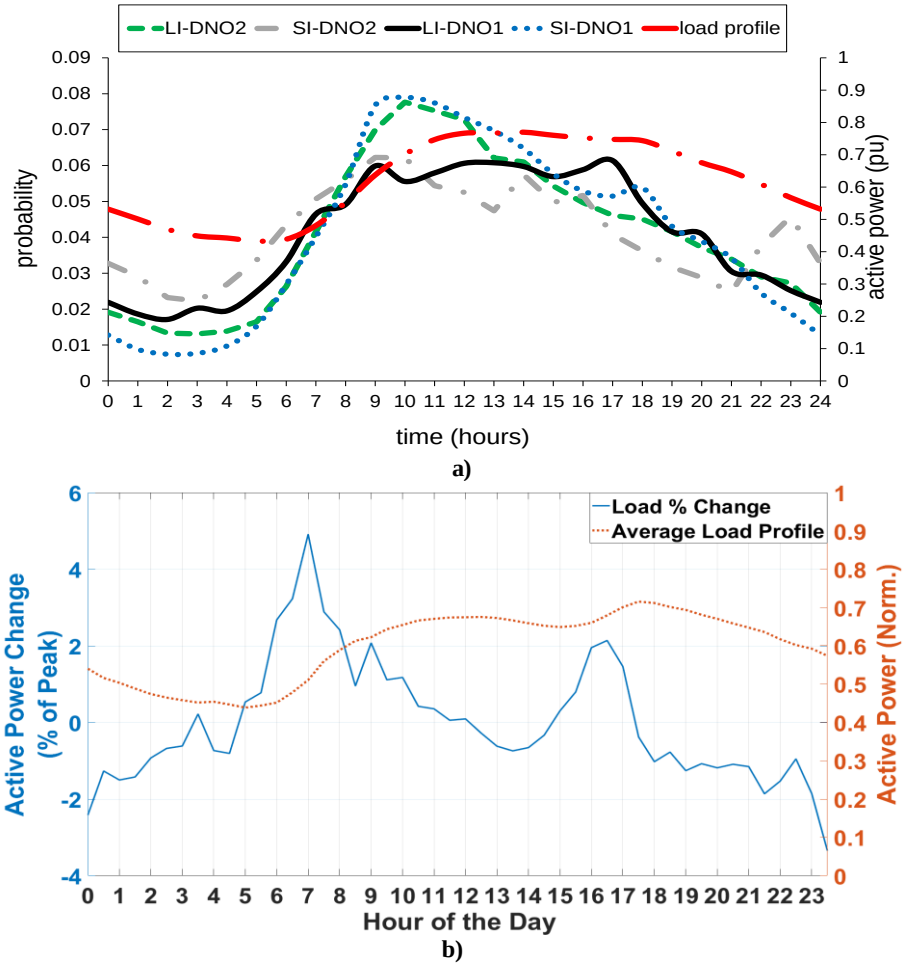


Figure 3.45: a) short-interruptions (SI) and long-interruptions (LI) form two European DNOs and b) rate of change of active power and average diurnal active power (form all GSP in the Scottish-A dataset)

Fault and interruption analysis is outside the scope of this thesis and any hypothesised correlations between the profiles of interruption-probability and rate of change of demands needs to be investigated further, particularly since the presented figures are restricted to the analysis of a limited number of both interruption and demand data.

3.9 Chapter Conclusions

The Fourier analysis has shown that the significance of components such as the: daily, half-daily, weekly and yearly cycles, for variables *P*, *Q* or *V*, is diverse and it varies for different GSPs, which opens the possibility for classification, as demonstrated by the GSP-groups in Tables 3.2, 3.3 and 3.4. For active power, daily and yearly cycles are the most significant modes of variability responsible for approximately 40 % and 33 % of the total range of variations, respectively (on average, based on 98 GSPs). The "reconstruction" of the active power demands, based on the first ten Fourier components (with the highest magnitudes)

performed particularly well, with correlation coefficients, above ~0.9, indicating that active power is determined by regular-predictable patterns, with lower penetration of stochastic components. Reactive power variability is shown to be determined primarily by the daily component (~37 %), the weekly component (~17 %) and to a lesser extent according to the seasonal/yearly cycle (~13 %), as determined through the analysis of reactive power measurements from 77 GSPs. "Reconstruction" of the reactive power signals performed poorer, compared with active power, with an average correlation coefficient of ~0.8 indicating less predictability, at least based on the analysis of inherent regular patterns (and not on regression with independent variables). For voltage, variations are generally restricted within limits of $\pm$ 6 % and these have been shown to vary according to the daily cycle (~15 %) and the yearly cycle (~ 15 %), which means that, from a statistical perspective, voltage variations can be considered (mostly) stochastic, as a large portion is attributed to higher frequency components. This is also reflected in the poor performance of the voltage signal "reconstruction", which is on average below ~0.7 (in terms of correlation coefficients), even with the inclusion of the first 100 Fourier components.

Weekly profiling (Section 3.4) has shown that there are significant differences in the demand levels for all three variables, between weekdays and weekends, as well as differences between pairs of days including Friday (both with weekdays and with weekends). These group difference have been established based on ANOVA and Kruskal-Wallis tests, which showed very good levels of agreement, for all variables and for (almost) all pairs of days of the week, with minimal exceptions. Active power demand differences between weekdays and weekends have been shown to be indicative of the percentage of total residential demand in individual GSPs (or similarly the % of domestic and non-domestic demands), as demonstrated in Figure 3.20. Periods of similar demand levels have been presented for the diurnal as well as the seasonal cycles in Sections 3.5 and 3.6. The specific diurnal periods in which the weekday/weekend differences are concentrated have also been presented. These profiles open the possibility of load identification as demonstrated in Figure 3.28, based on the assumption that these differences, for yearly-averaged profiles, are not attributed to weather related seasonal loads, but rather to loads related to occupancy levels and people's daily schedules.

The probability of occurrence of maximum and minimum demands, in Section 3.7, shows that these are concentrated during night hours, for minimum, and during mid-day and afternoon hours for maximum (which is assumed to be primarily determined by the customer-sector composition at each GSP). Despite the differences in demand patterns between weekdays and weekends, the periods of maximum and minimum demands are, almost, identical. Combined

diurnal and seasonal representation of the occurrence of maximum demands revealed patterns which are, potentially, associated with weather parameters (particularly solar irradiance). This is also shown Section 3.8, based on the rate of change of active and reactive power demands.

The profiling and normalisation methods presented in this chapter, are used in subsequent analysis throughout this thesis. In particular, the next chapter (Chapter 4), presents analysis of correlations and dependencies, based on seasonal and diurnal profiles, as well as on seasonal profiles on a per half-hour of the day basis.

Chapter 4: Correlations and Dependencies of Aggregate Demands

In the previous chapter, it has been demonstrated that parameters *P, Q* and *V,* exhibit variability with respect to the daily, weekly and seasonal scales. Distinctions have been shown for the significance of these temporal components among the three parameters, among groups of GSPs, as well as with respect to the overall presence of stochastic variations. Despite these differences, it has also been demonstrated that a relatively small number of FFT components was sufficient to reconstruct the original signals and particularly for active and reactive power demands, with average correlation coefficients of ~0.9 and ~0.8, respectively. A reasonable assumption that follows is that some common factors must exist, which affect electricity power demands. These factors are referred to as the "drivers of demand variability".

This chapter uses a number of statistical approaches (discussed in Section 4.3), to determine the main drivers of demand variability, the strength with which they affect electricity demand and the success of the various methods in identifying, evaluating and quantifying these relationships. The analysis is also aimed at determining variables that can be useful in disaggregation and forecasting studies, without the requirement that their variabilities are causally linked to electricity demand variability.

Sections 4.1 and 4.2 describe the meteorological and analemma (solar azimuth and elevation angle) variables used in the analysis. Sections 4.4 and 4.5 present the results with respect to the diurnal and the seasonal time-scales and Section 4.6 expands the seasonal analysis, on a per half-hour of the day basis. Section 4.7 presents an analysis of residuals, offers justification for the use multiple-regression in the subsequent chapters and discusses the use of linear and polynomial regression for modelling electricity demands. In Section 4.8, the use of a moving-window regression approach is presented, for the assessment of the sensitivities of demands to external conditions, for smaller subsets of the seasonal data.

Apart from the dependencies to exogenous variables, this chapter also presents results for the correlation of the electrical parameters with one-another. This is aimed to study their dependencies from a statistical perspective, in contrast with load modelling approaches that are generally concentrated on the instantaneous relationships between *P, Q* and *V*. The results and conclusions drawn from the analysis presented in this chapter are used for purposes of load identification, disaggregation and forecasting, in Chapters 5, 6 and 7.

4.1 Data Description

The available meteorological parameters presented in Table 4.1 are used as explanatory variables in the analysis of correlations and dependencies, as well as for demonstrating the corresponding methodologies discussed in the following sections of this chapter. These measurements have been obtained at the same geographical locations as the five MV-datasets presented in Chapter 3, Section 3.1, and for the same time periods. In cases where resolutions of measured data differed, the same resolution (e.g. 30-min. timestamps) has been adjusted.

Table 4.1: Summary of meteorological explanatory variables

Variables	Units	Corresponding MV-Dataset				
		Scottish - A	Scottish - B	English	Danish	Slovenian
Temperature	C^o	✓	✓	✓	✓	✓
Solar Irradiance	$W \cdot m^{-2}$	✓	✓	✗	✓	✓
Relative Humidity	%	✗	✓	✗	✗	✗
Atm. Pressure	mb	✗	✓	✗	✗	✗
Wind Speed	$m \cdot s^{-1}$	✗	✓	✗	✗	✗
Solar Azimuth Angle	^{o}	✓	✓	✓	✓	✓
Solar Elevation Angle	^{o}	✓	✓	✓	✓	✓

Temperature (in degrees Celsius - C^o) represents surface-air temperature, measured at a height of 1.25 to 2 meters from land surface and according to standardised procedures. Solar irradiance (in watts per meter squared - Wm^{-2}) consists of components of direct, diffused and ground reflected solar irradiance and it is therefore usually referred to as global solar irradiance. Relative humidity (in %) is defined as the percentage of actual vapour density with respect to the saturation vapour density (or pressure exerted by actual vapour to the saturation vapour pressure), where saturation levels are reached when air cannot hold any more water, at a given temperature. Atmospheric pressure (in millibars - mb) is the pressure exerted by the atmospheric weight due to the mass of air directly overhead the measuring apparatus, while wind speed (in meters per second - ms^{-1}) measures the motion of air mass with respect to the surface of the earth at a given location [191], [192].

The data from these five variables represents the mean values within the resolution windows, e.g. at a 30-minute resolution, temperature values correspond to the average temperature over the 30-minute period. Solar azimuth and solar elevation angles (in degrees - ^{o}) are also used and are defined in Section 4.2. Considerations of normalisation, standardisation and scaling, as outlined in Chapter 3, Section 3.2 apply to the meteorological and analemma variables as well.

4.2 The Solar Analemma Variables

Analemma diagrams are used to describe the apparent (relative) position between two celestial bodies from the point of reference of observers fixed at a certain position on one of them. Specifically, for the Sun-Earth system, the solar analemma diagrams mark the relative position of the Sun in the sky, as observed from a fixed position on the Earth's surface at constant time intervals, i.e. observations at a fixed time each day, throughout the course of (but not limited to) one calendar year [193]. The resulting observations are usually expressed in terms of horizontal and vertical deviations known as azimuth and elevation (or altitude) angles. The horizontal deviations are a result of the differences between true solar time and apparent local time, also referred to as the "equation of time", which is a function of the Earth's obliquity, i.e. axial tilt and orbital eccentricity, i.e. deviations from a perfectly circular orbit. The apparent vertical motion (elevation angle) is a result of the inclination of the Earth's axis relative to its plane of revolution around the Sun [193]. An example of an analemma diagram is shown in Figure 4.1, for Edinburgh city, UK, at 12:00 hours.

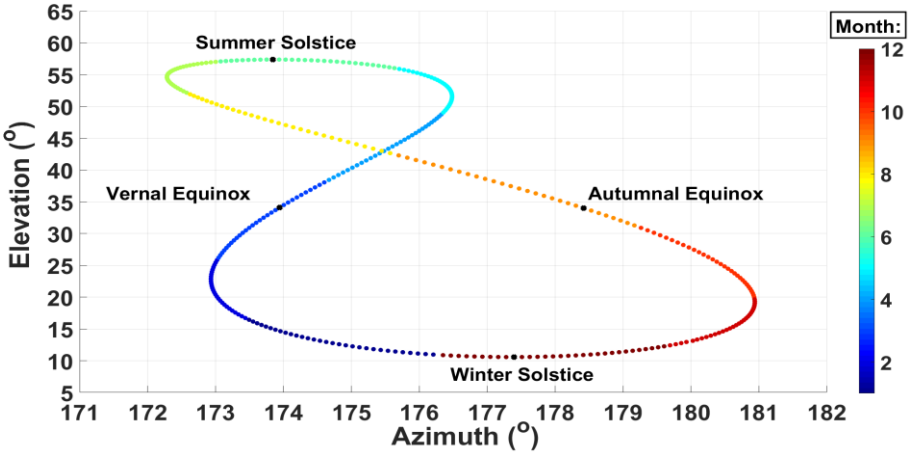


Figure 4.1: Analemma for Edinburgh, UK, at 12:00 hours

Marked on Figure 4.1 are the summer and winter solstices, i.e. the days of maximum and minimum excursion relative to the Earth's equator which correspond to the longest (summer) and shortest (winter) days of the year with respect to sunlight hours (in the Northern hemisphere - the reverse is true in the Southern hemisphere) and the vernal and autumnal equinoxes, when the centre of the Sun passes directly above the plane of the Earth's equator and correspond to days of, approximately, equal duration between night and day hours. Summer and winter solstices occur around the 21st of June and 21st of December, while vernal and autumnal equinoxes around the 21st of March and 21st of September. The exact dates however vary from one year to another.

The analemma variables are usually restricted to the diurnal periods when the Sun is directly observable in the sky (irrespective of atmospheric effects such as cloud coverage) and therefore during night hours the corresponding values are represented as constant zero measurements. This is compensated by the use of the "topocentric" azimuth and elevation angles, which allows the observed path of the Sun to extend to hours of the day when it is below the horizon, as shown in Figure 4.2, for all 48 half-hours of the day through the course of one year for Edinburgh, UK.

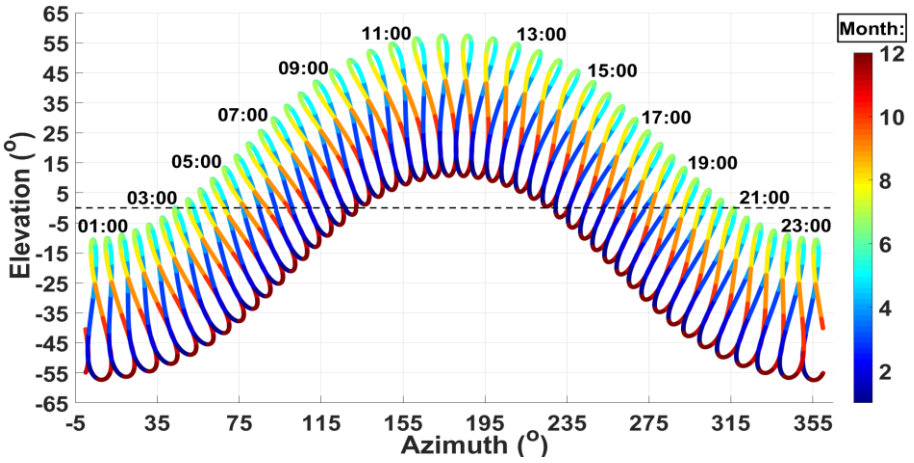


Figure 4.2: Topocentric analemma variables for one year, for Edinburgh, UK

Azimuth angles are set between 0º to 360º, with the 180º point at midday (12:00 hours), as in Figure 4.2, or otherwise between -180º to 180º with the midpoint at 0º. The horizon can be, conventionally, set at 0º (dashed horizontal line) in the elevation axis, but in reality, sunlight periods extend beyond this point due to atmospheric refraction effects. A more detailed description and information regarding the calculation of the analemma variables can be found in [193], [194] and [195].

Analemma diagrams are primarily used in astronomy, but knowledge of the relative position of the Sun in the sky has applications that extend to meteorology, climatology and agriculture as well as in the energy industry and particularly for solar energy (e.g. Sun tracking) technologies, as discussed in Chapter 2. Within the context of electricity supply and demand interactions, these variables can offer new approaches for the analysis of seasonal dependencies that have further applications for load profiling, demand forecasting and load identification and disaggregation studies. The relative position of the Sun in the sky has the benefit of simultaneously capturing seasonal and diurnal cycles, while it is also directly related to weather variables such as solar irradiance and temperature levels. However, unlike weather conditions, the solar analemma is unaffected by atmospheric phenomena. The underlying processes are relatively stable and deterministic (at least for up to several decades) and the

analemma variables can be calculated to a high precision for any location on the Earth's surface and for any desired period of time. This is particularly useful for the analysis of dependencies in demand variability in the absence of meteorological variables. In previous publications, e.g. [89], some variables closely related to the solar analemma variables have been used, such as the sunrise and sunset times, daylight-hours and daylight-duration variables, but to the best of the author's knowledge the solar analemma variables have not been used for the assessment of electricity demand variations in the manner presented in this thesis (by the author: [67]).

4.3 Overview of Applied Statistical Approaches

4.3.1 Definitions

In the context of statistical analysis, the terms correlation and dependence are used to describe the existence and when quantified, the extent of associativity between two random variables or, more specifically, the extent to which these variables change (fluctuate) together. In the same context, a random variable does not refer to a variable whose function (or procedure) is non-deterministic, but rather to the fact that the possible values acquired by the function are defined by some probability distribution and then mapped onto a state-space (e.g. real numbers), such that:

$$X: \Omega \rightarrow E \qquad (4.1)$$

where X is a random variable from a probability space Ω, mapped to a set in measurable space E. For example, active power demand (e.g. from a single GSP) can take any value within a particular range of a distribution, with the probability given by the integral of the probability density function, within the prescribed range. By determining the associations between dependent (e.g. active power) and independent/explanatory (e.g. temperature) variables, correlation analysis (and specifically regression analysis when a fitting function is calculated) aims to limit the range of the probability density function, so that the values of the dependent variable can be more accurately predicted from the known values of the independent variable(s). Knowledge of these associations also allows for better understanding of the underlying factors that determine the changes of the dependent variable (i.e. explained variance) and can therefore be used to draw conclusions about its characteristics, composition and time-dependent evolution.

While the existence of associations between variables implies causal relationships, it cannot demonstrate or prove them. This problem is a well-documented one, as it penetrates all forms of scientific inquiry and it is therefore studied by various disciplines such as statistics, physics,

engineering, experimental design and medical studies, while the concepts of proof and proof of causality are also discussed in branches of philosophy such as epistemology. Although it is not within the scope of this thesis to study or discuss these issues in depth, it should be noted that there are some minimal conditions which, when met, can establish causal relationships within some acceptable limits of certainty. These include: a) the strength of the associations, which are quantified using various correlation coefficients, b) the consistency of the relationship among various data samples (or experiments), as well as its reproducibility and perhaps most importantly, c) the development, or support of an underlying theory that can explain an observed relationship.

4.3.2 Metrics of Correlation and Regression Analysis

The strength of the correlations between samples of two variables, x and y, can be quantified using the Pearson's product-moment correlation coefficient (also used in Chapter 3), defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (4.2)$$

where the numerator denotes the covariance between the two samples and the denominator the product of their standard deviations [196]. The use of the Pearson's coefficient is based on assumptions of a linear relationship between the two variables, as well as the use of continuous variables. The first can be overlooked, but if violated, it will inevitably affect the output coefficient, i.e. a perfectly dependent non-linear relationship will have a lower correlation coefficient because the estimation cannot account for the non-linearity. The correlations can also be quantified using the Spearman's rank correlation coefficient, defined as:

$$r'_{xy} = \frac{cov(rg_x, rg_y)}{\sigma_{rgx}\sigma_{rgy}} \quad (4.3)$$

which is essentially the Pearson's correlation coefficient as applied to ranked variable values and while the Pearson's analysis assesses linear relationships, the Spearman's coefficient can account for monotonically increasing or decreasing relationships, which are not necessarily linear [197]. In both cases, the results are presented using coefficients within the range of [-1 1], where -1 denotes a "perfectly" negative correlation and +1 a "perfectly" positive correlation. Furthermore, and in both cases, no prior assumptions are made for the distinction between the two variables being dependent or independent, nor does the analysis provide information about the direction of causation (if any) between the two variables. Other correlation coefficients can be found in literature, with their applicability depending on the specific analysis and type of

input data samples. For the analysis presented in this chapter, the Pearson's and Spearman's correlation coefficients are considered.

A more involved description of a relationship between two variables is possible with the introduction of a "best fit" equation that, in the case of simple linear regression, is in the form of:

$$y = \beta_1 x + \beta_0 \tag{4.4}$$

where $\beta_1$is the gradient of the linear function, β_0 is the y-intercept and y and x are the dependent and independent variables, respectively. In most situations the "best fit" is an approximation (not based on a perfect correlation) and therefore, for the actual values of y, i.e. y_i, the solution is in the form of:

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \tag{4.5}$$

with the distinction being the introduction of an error term, or residual, ε_i, for each data-point i. The term "simple", in simple linear regression, corresponds to the use of only one independent or explanatory variable, x, while it is possible, as in the case of multiple linear regression, to predict the dependent variable based on the variations of two, or more, explanatory variables. The term "linear", in the same context, does not refer to the fact that the equation is of the 1st polynomial, i.e. x^1 order, but rather to the linear mapping between the estimated parameters and the outcome, i.e. βx and not x^β.

The most commonly used approach for estimating these parameters is referred to as ordinary least squares (OLS) method, which achieves fitting by the minimization of the sum of squared residuals, between the predicted and actual values of y, [198]. The OLS method can be used for polynomial fitting of a higher degree, such as 2nd and 3rd, as well as for multiple regression with more than one explanatory variable (both discussed in Section 4.7). For the OLS algorithm, the two are actually equivalent and polynomial regression can be considered a special case of multiple regression [199].

A main problem that arises from higher degree polynomials, as well as in the case of multiple regression models, is that it complicates the interpretation of the, multiple, resulting regression coefficients. Moreover, the "appropriation" of variance to the various independent variables is also problematic in the presence of multicollinearity (discussed in more detail in Section 4.3.3). Higher degree polynomials are particularly useful when analysing non-linear relationships. An example would be the analysis of active power demands for temperature variations, at a geographical location where there is significant penetration of thermal heating loads during the winter period and AC/HVAC loads during the summer period. The simple

linear approach is not suitable for capturing this relationship, unless two data samples are created based on temperature turning-points, discussed in Chapter 6. When dealing with demands demonstrating such effects, the approaches presented in the following sections should be appropriately modified to account for the non-linear dependencies.

The goodness-of-fit of the linear regression models is assessed by calculating the portion of the initial variance in the dependent variable that can be explained (or predicted) by the independent variable:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}} \qquad (4.6)$$

which is referred to as the coefficient of determination – R^2 [200]. In the case of simple linear regression and when the y-intercept is included, as in (4.4), the coefficient of determination is equal to the squared Pearson's correlation coefficient.

4.3.3 Multicollinearity

Multicollinearity is a common problem associated with regression analysis, particularly when the analysis is aimed at determining the relative strength of the effect of each explanatory variable to the dependent variable. Multicollinearity describes the situations in which explanatory variables are themselves strongly correlated with one-another. While this does not necessarily affect the predictive power of a model, such as in multiple regression forecasting, the effects of each individual variable are harder to decipher. This is better illustrated in Figure 4.3 using the results from factor analysis, for a single GSP (GSP-14) at 17:00 hours, for the duration of one-year and considering: active power, reactive power, temperature, solar irradiance and solar elevation angle.

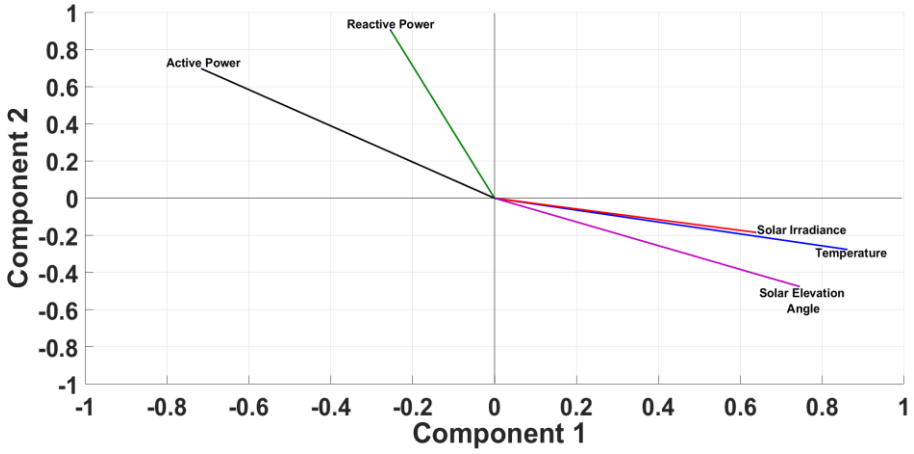


Figure 4.3: Example of factor-analysis results for 5 variables at 17:00 hours, GSP-14

Factor analysis is a procedure which aims to reduce the dimensionality of the variability of observed parameters in a way that these can be expressed in terms of reduced components, i.e. factors, or "latent" factors [201]. The vector sizes, or loadings, of each variable correspond to how much of their variance can be explained by each factor, i.e. component. More importantly, for this discussion, the angle between the variables indicates their respective collinearity or covariance. For example, perfectly orthogonal variables have variations that are shown to be independent of one another. The example presented above is restricted to two components but more can be added to allow for more explained variance. As it is shown in Figure 4.3, groups of variables change together, such as: temperature, solar irradiance and solar elevation angle, or active power and reactive power. Therefore, when examining the effect of e.g. solar irradiance on active power demand, it is possible that some of the explained variance may actually be attributed to the correlated changes of temperature with solar irradiance. This is particularly important when the analysis is aimed at determining portions of active power demand that can be attributed to specific load-types, such as lighting loads and heating/cooling loads and these effects are therefore further discussed in Chapter 6. More information about factor analysis and related methods such as principal component analysis (PCA) can be found in [201], [202] and [203].

Partial correlation analysis can be used to account for multicollinearity effects, by determining the strength of associations when the effects of the rest of the independent variables are controlled [204]. A comparison between the explained variance (in terms of squared correlation coefficients - r^2), between zero-order (regular) correlation (4.2) and partial-correlation is presented in Table 4.2, considering active power as the dependent variable and temperature and solar irradiance as the independent variables; for a single GSP (GSP-14) at 17:00 hours, for the duration of one-year.

Table 4.2: Example of zero-order and partial correlation results

Independent Variables	Zero Order r^2	Partial r^2
Temperature	0.65	0.49
Solar Irradiance	0.34	0.03

The squared correlation coefficients drop for both explanatory variables in the case of partial correlation analysis, due to the high degree of multicollinearity (demonstrated in Figure 4.3). If, in the above example, active power is represented by variable *y*, temperature by *x* and solar irradiance by *z*, then the partial correlation coefficient between active power and temperature is given by:

$$r_{yx.z} = \frac{r_{yx} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}} \qquad (4.7)$$

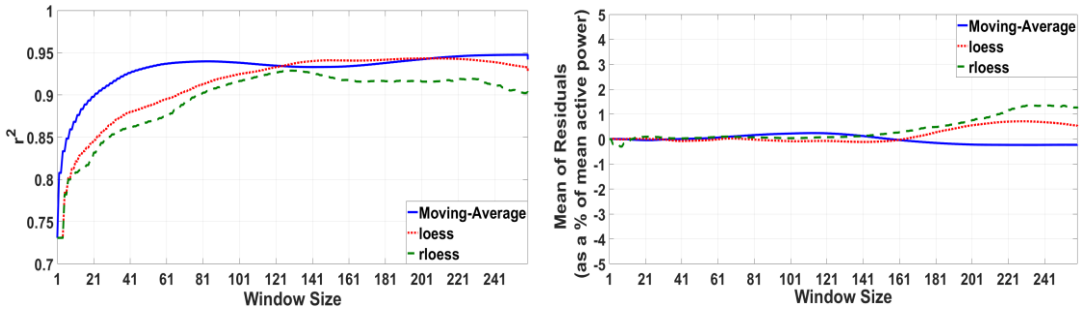
However, criticism can be found regarding the efficiency and validity of partial-correlation analysis, particularly in cases where the multivariate normality is violated, or when effects of non-linearity in the relationships are present [205], [206]. Furthermore, the results from partial correlation analysis vary according to the selected inputs, i.e. explanatory variables. For example, the inclusion of solar elevation angle in the above example (Table 4.2), would results in a decrease in the partial r^2 values of both temperature and solar irradiance. Partial correlation results are therefore particularly sensitive to model specifications and the correlations between a dependent and an independent variable vary according to the number of selected control variables. The analysis presented in the following sections is based on the regular (zero-order) correlations because the aim is to assess the common variability that can also be used for predictive analysis (such as forecasting in Chapter 7). For regression, determining partial correlations involves the analysis of residuals and more specifically the correlations of residuals from the different pairs of dependent-independent and independent-independent variables. Discussion about the analysis of residuals, their periodicities, autocorrelations and how conclusions from such results can be used to expand the methodology into multiple and polynomial regression models is provided in Section 4.7, while more involved approaches for specific "appropriation" of variance, in the context of load disaggregation, are discussed in Chapter 6.

4.3.4 Filtering of Samples

Regarding the pre-processing phase of the analysis, the input samples can be filtered, i.e. smoothed, in order to be refined, e.g. by reducing "irrelevant" high-frequency variations, so as to highlight certain frequency components and therefore increase the correlations between demand data and exogenous parameters.

The effects of data-smoothing on the correlations of active power with temperature are presented in Figure 4.4 (a), in terms of the resulting correlation coefficients (Pearson's), for data corresponding to one-calendar year (weekdays only), for a single GSP, at 17:00 hours and for increasing filtering window lengths. Filtering, in this case, is applied to reveal the seasonal components, by reducing the day-to-day fluctuations. A moving-average filter is used, as presented in Chapter 3 (3.13), as well as the "loess" and "rloess" filters, which correspond to: a local regression using weighted least squares and a 2nd degree polynomial model; and a robust version of "loess" that assigns lower weights to outliers in the local regression model, [207] and [208]. The correlations rapidly improve for the moving-average filter, for up to ~60 days of window length, at which point a saturation level is reached and no particular change

in the coefficients can be shown. Saturation levels are reached for the "loess" and "rloess" methods as well, but these are found at a later stage, i.e. around 130 days of window length.



a) b)

Figure 4.4: a) squared Pearson's coefficients for active power with temperature, for increasing window lengths of filtering of both variables and b) mean value of the active power residuals (as a % of mean active power) for increasing window-lengths of data filtering

The suitability of an adopted filtering approach can be investigated by considering the distribution of the residuals, defined, in this case, as the differences between the actual and smoothed values of the input variable(s). An example is presented in Figure 4.4 (b), showing the mean of the residuals of active power, for the three methods discussed above and for the full range of window lengths (for the same GSP and half-hour of the day as in (a)). The mean values of the residuals do not exceed a ±2 % threshold, with respect to the mean value of the actual active power inputs, and it is generally kept very close to zero for all considered window lengths. The slight positive bias shown for window lengths of more than ~160 days is due to the fact that, for a large window size, the filtering procedure determines the annual trend instead of removing the high-frequency random fluctuations and can be therefore considered as an inappropriate pre-processing approach when the aim is to capture the seasonal covariance of two parameters. It is generally desired that the mean value of the residuals is kept as close to zero as possible in order to reduce the bias of the filter towards positive/negative errors and to ensure that the filter removes only what can be considered as random Gaussian noise (Kolmogorov-Smirnov and Jarque-Bera tests of normality can also be used). In the same context, it is also important to ensure that the residuals have low levels of autocorrelations, so that it can be shown that the removed fluctuations do not include non-stochastic, periodic components (similar to the analysis of residuals for the regression models, presented in Section 4.7). Despite the fact that filtering can improve the correlations between pairs of variables, some important issues need to be taken into consideration, to avoid reaching erroneous conclusions. Firstly, data filtering can result in correlated samples that are actually non-dependent, simply by exhibiting long/short term components that, by chance, are found for both random variables. The phenomenon of spurious correlations is not limited to cases when filtering is applied, but the high correlation coefficients that often result from smoothed

datasets can increase the chances of false-detection of correlations. Furthermore, if the application of smoothing has the general tendency of increasing the correlation coefficients for a given pair of variables (such as active power and temperature), the results (for the smoothed data) do not offer any new information since these improvements are expected to be common for all input data-samples, for which initial correlations can be found. Similarly, when conducting regression analysis, the increased coefficients of determination - R^2 can be misrepresentative of the actual predictive/forecasting power of the models. The problem can be thought of as a case of "overfitting", in which artificially highly correlated samples are created.

There are, however, cases where filtering is justified by the fact that, e.g. strong variability in high frequency components can actually mask overall seasonality and such cases are also investigated, as in Section 4.6. For the rest of the presented analysis, the input datasets are non-filtered and when filtering is applied, justification is provided as well as discussions based on the corresponding results.

4.3.5 Moving-Window Regression

Moving-window (or rolling-window) regression is an approach by which regression coefficients are calculate for smaller samples of the initial input data, corresponding to the data-points within a desired window length, thus assessing the correlations within a smaller time-frame. The final representation, for the complete duration of the input data, can be thought of as an alternative to determining the coefficients for various seasons of the year and instead the analysis can produce a continuous set of coefficients, thus also assessing the sensitivities of the dependent variable to the independent variable(s), throughout the year. The method is therefore a combination of regression analysis (4.4) and of a moving-window filter (3.13), but modified to evaluate linear regression coefficients. An example is the computation of the coefficients between active power and temperature, at a particular half-hour of the day, through the course of one calendar year, as presented for four different window-lengths, i.e. for 10, 20 ,40 and 60 days, in Figure 4.5.

The coefficient of determination - R^2 (4.6), shown in (a), as well as the gradient of the linear fit - b (4.5), shown in (b), tend to stabilise and present a clearer seasonal pattern for moving-windows of length more than 20-days, i.e. better results for 40 and 60 days. The data samples are simply too small to account for any actual relationships for smaller window-lengths and therefore the resulting coefficients are "unstable" as shown by the more pronounced day-to-day fluctuations, for 10-days and 20-days. The trade-off for longer window sizes is the reduction in resolution (i.e. finer detail dependencies are captured by smaller regression

windows), while the selection should also be based on an inspection of the resulting R^2 values, so that the gradients can be accompanied by high/higher statistical significance.

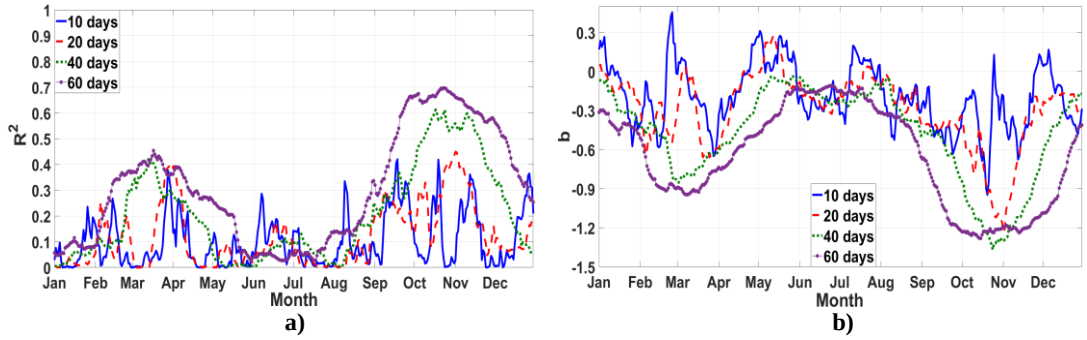


Figure 4.5: Moving-window regression results for active power with temperature at 17:00 over one year, for four different window-lengths (10, 20, 40 and 60 days)

Similar to the use of data-filtering (Section 4.3.4), the selection of an appropriate window size for the moving-widow regression is not straightforward and does not rely on some well-defined principles. In general, the decision depends on the particular application of the analysis and the characteristics of the input datasets and no universal agreed window size exists. In this chapter and for the moving-window regression analysis presented in Section 4.8, a window length of 41 days is used, which corresponds for ± 20 days on either side of each data-point, i.e. a 2-month period. This has been shown to be sensitive enough to determine the relative changes in the gradient of the dependencies, while also considering a sufficiently large data sample do ensure statistical validity.

4.3.6 Further Considerations

The results presented in the following sections are often expressed in terms of percentile values, i.e. median, 95th and 5th percentiles. The choice for this representation is based on the fact that the inclusion of a large number of input samples produces coefficients that have (in some cases) high variability among the various GSPs and do not always form normally distributed sets. Therefore, non-parametric statistics are more appropriate, since parametric statistics can often be misleading due to the skewness of distributions, or the presence of outliers, failing to show measures of central tendency. Furthermore, the use of the 95th and 5th percentiles is an intuitive representation of the range of the resulting coefficients.

Regarding considerations of statistical significance and following the discussion provided in Chapter 3, Section 3.4, the current analysis is limited to the presentation of correlation coefficients (Pearson's and Spearman's) and of coefficients of determination - R^2. The use of p-values, as well as the use of confidence intervals (CIs) for the resulting beta coefficients (in regression analysis), are both omitted for purposes of efficient and clearer presentation of the

results. Confidence intervals are assigned for a specific confidence level (usually at 95%) and are measures of uncertainty associated with the sampling technique, i.e. within the computed intervals, for an arbitrary number of samples, the parameters estimated by the model would be within the newly estimated parameters for 95 % of the samples. This means that the resulting coefficients and the percentile values (median, 95th and 5th) should be estimated for the upper and lower bounds of the corresponding confidence intervals and the resulting figures would require many more curves for complete representation. These metrics are of course important components of any detailed statistical description, particularly when the results are then used for further analysis but, for the purposes of presenting the general tendencies of the correlations between the considered variables these are deemed redundant, in the sense that their interpretation is almost impossible when they are averaged over all data samples. However, when the methodologies are applied to specific GSPs to study relationships between specific variables, these metrics can be added to increase the overall validity of the results.

4.4 Diurnal Correlations

Diurnal correlations measure the degree of associations between active power, reactive power and voltage as well as between these three variables and the meteorological and analemma parameters (Sections 4.1 and 4.2) over the diurnal period, i.e. 48 data-points, in the case of half-hourly measurements. The analysis is conducted for each day individually and the results are presented for the duration of one calendar year. The methodology can be summarised in the following steps:

- For each variable, values are normalised with respect to the maximum in each dataset (3.3). For each pair of variables, the diurnal correlations are quantified using the Pearson's and Spearman's correlation coefficients (Section 4.3). As no substantial differences can be reported between the results of these two, only the Pearson's coefficients are presented[7].
- The resulting datasets include 365 coefficients (one per day), for each pair of variables and for all GSPs, depending on data availability[8]. The 50th (median), 95th and 5th percentile values are then calculated using the results from all available GSPs. The arrays of

[7] The relationships do not deviate from the linear and the rank approach (Spearman's) gives, approximately, the same coefficients as the Pearson's. No consistent differences or differences greater than $r = \pm 0.1$ can be reported.

[8] The number of GSPs available for the *P*, *Q* and *V* parameters, are as presented in Chapter 3, Table 3.1 and the data availability for the weather and analemma variables is as presented in Table 4.1.

coefficients are further processed using a moving-average window filter (Section 4.3) to allow for clearer presentation of the yearly trends.

The relative distances between the median, 95th percentile maximum and 5th percentile minimum values are indicative of the consistency of the correlations over all GSPs used, for each pair of variables, as well as of the consistency with respect to the day-to-day fluctuations of the resulting coefficients within the yearly period. As a consequence, the results are not descriptive of individual GSPs, but rather of the overall characteristic relationships between the selected variables, in the diurnal time-frame. The seasonality of the percentile metrics shows whether the diurnal correlations exhibit noticeable changes throughout the year. All of the above are better illustrated using the results presented in Figure 4.6 for: a) active power with reactive power, b) active power with voltage and c) reactive power with voltage.

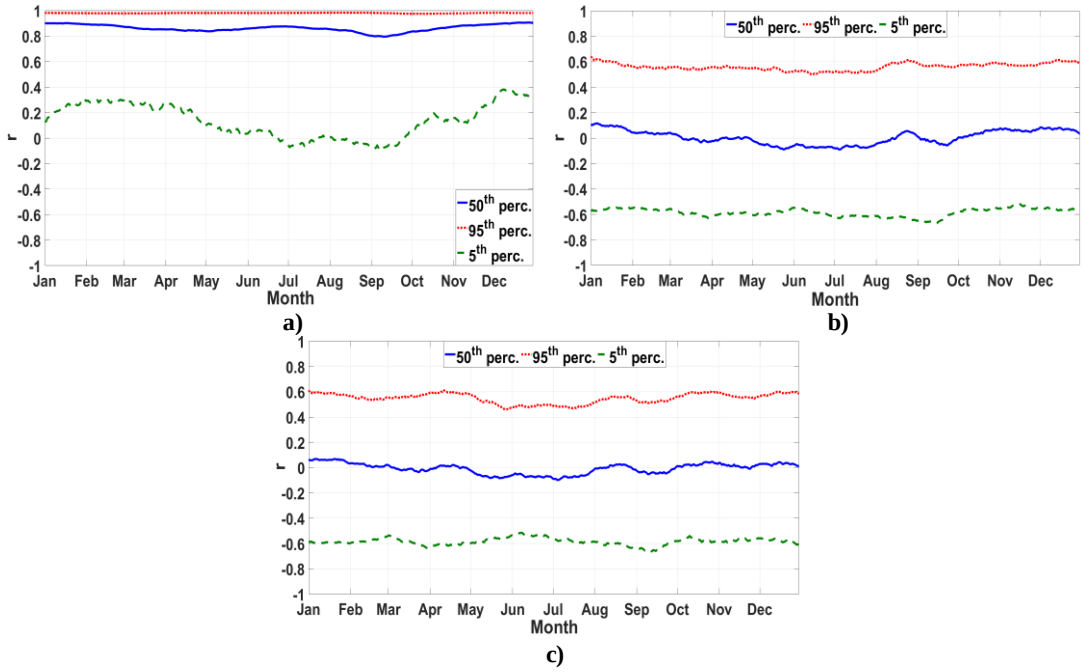


Figure 4.6: Diurnal correlations over a one-year period for: a) active power with reactive power, b) active power with voltage and c) reactive power with voltage

For active power and reactive power, the median and 95th maximum values are both over 0.8 indicating that, generally, there are very strong positive correlations between active power and reactive power over the daily cycle, with no apparent seasonal changes and with high consistency among the various GSPs. However, as it can be seen from the 5th percentile minimum values, there are weaker correlations for some of the GSPs. In particular, the negative correlations correspond to GSPs-19 and 43, which have also been discussed in Chapter 3, for showing atypical demand cycles, for both *P* and *Q*. GSP-19 was also grouped in a single-GSP group in the Fourier analysis in Section 3.3.

For active power with voltage, in (b), as well as for reactive power and voltage, in (c), the median values are close to zero with the range of maximum and minimum values equally spaced around the median, indicating weak correlations for the two pairs of variables. The results demonstrate that, over the diurnal cycles, the variations of voltage levels cannot be statistically associated with the variations of either active power or reactive power (based on the samples used in this analysis, which are limited to the 24 English GSPs for which voltage measurements were available). These results are also supported by the findings of the Fourier analysis, Chapter 3, Section 3.3, which demonstrated that voltage variations are mostly stochastic and that the daily component cannot account for more 15 % (on average) of the total range of variations.

Figure 4.7 shows the results for the diurnal correlations between active power and meteorological variables: a) temperature, b) solar irradiance and c) relative humidity.

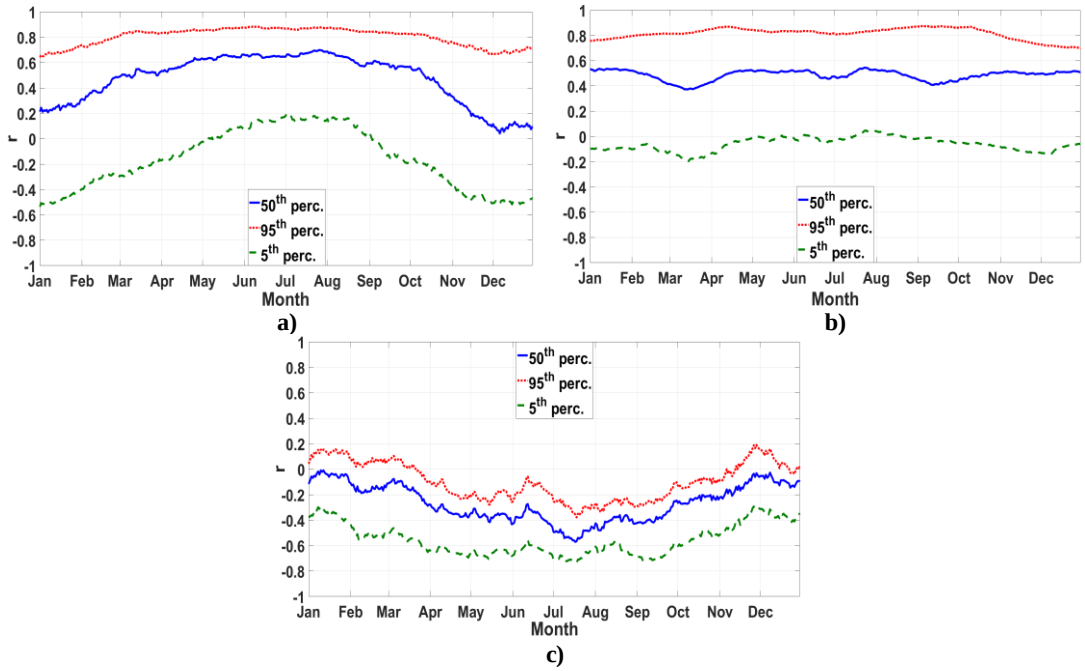


Figure 4.7: Diurnal correlations over a one-year period for: a) active power with temperature, b) active power with solar irradiance and c) active power with relative humidity

For active power with temperature, Figure 4.47 (a), the correlations are positive for the median and 95th percentile values while there is an apparent seasonal trend, with increasing strength of correlations for the summer period. The relationship is the reverse of what is expected for (considerably) colder climates, where a decrease in temperature is usually associated with an increase in active power due to increased demands for heating loads. This is also the case for the correlations with solar irradiance, in Figure 4.47 (b), which are at a level of ~0.5, for median values, with a range between -0.2 to 0.9 and with no seasonality trend shown. Relative

humidity, in Figure 4.7 (c), has negative correlations with active power, which tend to increase in strength during the summer months, a relationship that is the reverse of the one observed for active power with temperature. This is because as air gets warmer it can hold more water vapour and thus its saturation levels increase which means that relative humidity decreases (definition of relative humidity in Section 4.1 and further discussion in Section 4.6). The effects become more pronounced during the summer months and particularly for the daily periods when temperatures reach their maximum for the day. This is also the reason why correlations between active power and temperature are stronger during the summer periods, i.e. greater difference between night and day temperatures means that the diurnal temperature cycles are more pronounced and thus better correlated with the diurnal demand cycles. No statistically significant[9] correlations can be reported for active power with atmospheric pressure, as well as for active power with wind speed and therefore these results are not presented.

Figure 4.8 shows the results for active power with the analemma variables: with solar azimuth angle in (a) and with solar elevation angle in (b). The correlation coefficients are positive apart from the 5th percentile for elevation angles and between 0.4 to 0.7 for median values (higher and more consistent for active power with elevation angles).

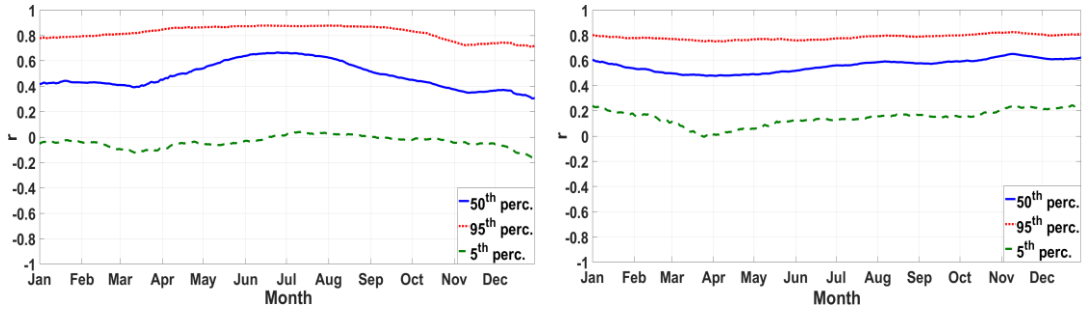


Figure 4.8: Diurnal correlations over a one-year period for: a) active power with solar azimuth angle and b) active power with solar elevation angle

Despite the fact that strong correlations are shown between active power and temperature, solar irradiance and azimuth/elevation angles (Figures 4.7 and 4.8), these cannot be directly associated with changes in loads affected by the corresponding explanatory variables. The reason can be explained with reference to Figure 4.9. Active and reactive power demands have similar diurnal "paths", which are strongly and positively correlated with the remaining

[9] This refers to the values of the correlation coefficients. Although there is no clearly defined threshold and statistical significance would require further analysis; an *r* value below ± 0.3 is generally regarded as very weak to no correlation. Furthermore, since these low values were consistent for all GSPs, the corresponding pairs of variables are not presented.

variables (excluding relative humidity). Azimuth angle has an increasing trend throughout the day due to the constant east to west motion of the sun, which correlates with the increasing trends of both active power and reactive power, which reach a peak during mid-day and afternoon hours. Solar irradiance and solar elevation angles follow a similar diurnal pattern and peak during the mid-day period (although solar irradiance does not always peak during that period due to cloud coverage). Temperature has a daily distribution with a peak lagging the solar irradiance and elevation angle peaks, which can be explained by the time required to heat the earth's surface and atmosphere.

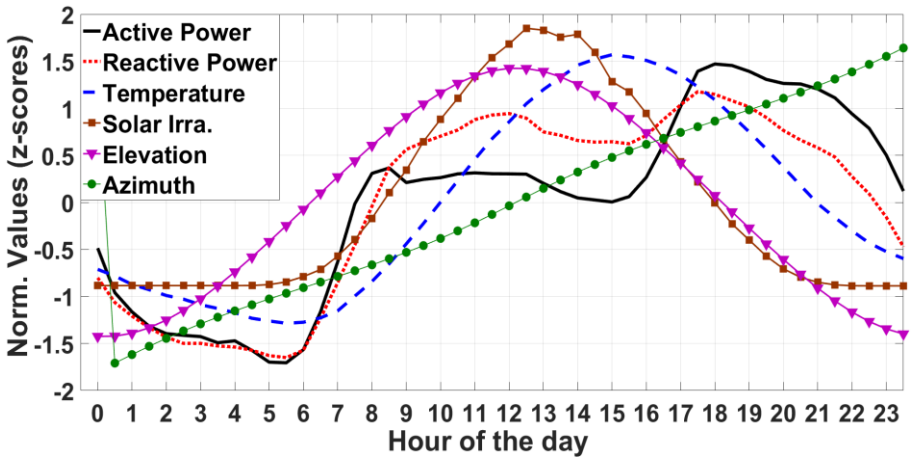


Figure 4.9: Three-year average of diurnal profiles for selected variables

In the diurnal time-frame, solar azimuth and solar elevation angles are determined exclusively by the earth's rotation, but essentially the same motion is responsible for solar irradiance levels, as well as for people's daily schedules (and even more so based on the "working-hour schedule"), which is represented by the diurnal patterns of active and reactive power demands. Therefore, while these diurnal patterns are functions of the same underlying processes (i.e. alternation of day and night, as a result of Earth's rotation), they do not necessarily have a causal relationship in the sense that, e.g. active power demand does not increase because temperature or solar irradiance increase. In reality, the reverse is usually true, at least for temperate/cold climates, as it is shown in the correlation analysis from a seasonal perspective, in the following sections (4.5 and 4.6). The only pair of variables for which the correlations cannot be regarded as purely consequential of the diurnal cycle are active power and reactive power, because these are causally linked by the increasing demands for loads that require both *P* and *Q* and the correlations are therefore a result of electrical characteristics.

Because of the strong positive correlations between active power and reactive power, Figure 4.6 (a), the correlations between reactive power and temperature, solar irradiance, relative humidity and solar azimuth and elevation angles, are at similar levels, with similar yearly

patterns as with active power demand. Furthermore, and following the discussion provided for Figure 4.9, i.e. non-causally linked correlations, the results for reactive power with the meteorological and analemma parameters are not presented[10]. Regarding voltage, no statistically significant correlations[11] can be reported with any of the meteorological or analemma variables, in the diurnal cycle.

The results presented in Chapter 3, Section 3.3 show that, on average for all available GSPs, diurnal variations can account for ~41 % (± 13 %) and ~36 % (±17 %) of demand variability for active and reactive power, respectively. The analysis presented in the current section indicates that, while correlations between demands and meteorological/analemma variables exist in the diurnal time-frame, these cannot be causally linked and thus it can be concluded that diurnal variability is regulated, primarily, by people's daily schedules and the generally constant consumption patterns, at the corresponding GSPs. These conclusions do not imply that changes in the external parameters have absolutely no effect on electricity demand within the daily period. What it is rather shown, is that the diurnal demand profiles, or diurnal consumption patterns, are relatively fixed and that deviations of demand levels due to weather parameters do not, significantly, shift the diurnal patterns to a degree that would make these changes detectable based on a diurnal correlation analysis. In order to capture and quantify the sensitivities to these external conditions, it is necessary to consider the seasonal correlations (Section 4.5), the seasonal correlations on per half-hour of the day basis (Section 4.6), as well as the seasonal correlations for data sub-sets based on a moving-window regression analysis (Section 4.8).

4.5 Seasonal Correlations

The seasonal correlations between pairs of variables corresponds to the level of their associations throughout the course of one calendar year. These correlations are evaluated based on the daily-average, daily-maximum and daily-minimum values, between active power, reactive power and voltage, as well as between these three parameters and the meteorological and analemma variables presented in Sections 4.1 and 4.2. Weekdays only are considered, to minimize the effects of having input samples with two distinctive levels of

[10] These are, however, almost identical to the results presented for active power in Figure 4.7, with slightly higher correlations between reactive power and solar irradiance and elevation angles, due to the fact that reactive power more frequently peaks during the mid-day period (compared with active power), resulting in better matching with the corresponding diurnal patterns of the two variables.

[11] These are all at levels below $r = \pm 0.2$ and no consistent correlations are shown among the 24 GSPs used for the analysis.

demands, i.e. due to the differences between weekdays and weekends, as demonstrated in Chapter 3, Section 3.4. The approach can be summarised in the following steps:

- Average-daily values, maximum-daily values and minimum-daily values are calculated for active power, reactive power and voltage, for each of the 261 weekdays of the year and for all available GSPs. The corresponding values are calculated for the independent variables (i.e. weather and analemma) and with respect to data availability for each MV-dataset, as discussed in Section 4.1. In all cases, values normalised by (3.3) are used.
- The seasonal correlations are then evaluated, for the three daily statistics (mean, max. and min.), for each combination of variables and for all available GSPs. The results are quantified using the Pearson's and the Spearman's coefficients.
- Then, for each pair of variables the results from all available GSPs are summarised using the 50th percentile (median), 95th percentile and 5th percentile values.

The analysis has shown that there are no consistent differences between the correlation coefficients, when using the average-daily values, the maximum-daily values or the minimum-daily values as inputs. This is demonstrated in Figure 4.10, based on the analysis of the seasonal correlations of active power with reactive power.

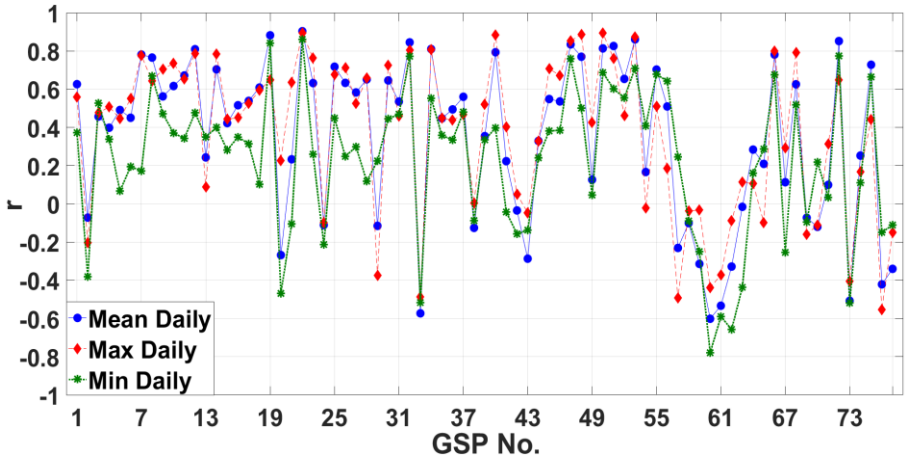


Figure 4.10: An example of correlations based on the mean-daily, max.-daily and min.-daily values, for active power with reactive power, for 77 GSPs

Although, in general, lower coefficients are obtained for the minimum-daily values, for ~ 50 % of the GSPs, the differences among the three daily statistics are not consistently higher or lower for the whole dataset and maximised correlations can be shown based on the analysis of all three. Therefore, for the available data, seasonal correlation of the daily extremes (i.e. max. and min.) does not produce results that substantially differ from the results obtained using the average values. Furthermore, using the Spearman's coefficient, as opposed to the Pearson's

coefficient, also does not produce significantly different results. An example is shown in Figure 4.11, based on the seasonal correlations of active power with reactive power (using average-daily values). It can be concluded that, on average, for all considered datasets, the seasonal relationships do not substantially deviate from the linear, although it can be shown that for individual GSPs and pairs of variables, non-linear regression can improve the correlation results and better estimate the underlying relationships. This is discussed in more detail in Sections 4.3, Section 4.7 and in Chapters 6 and 7.

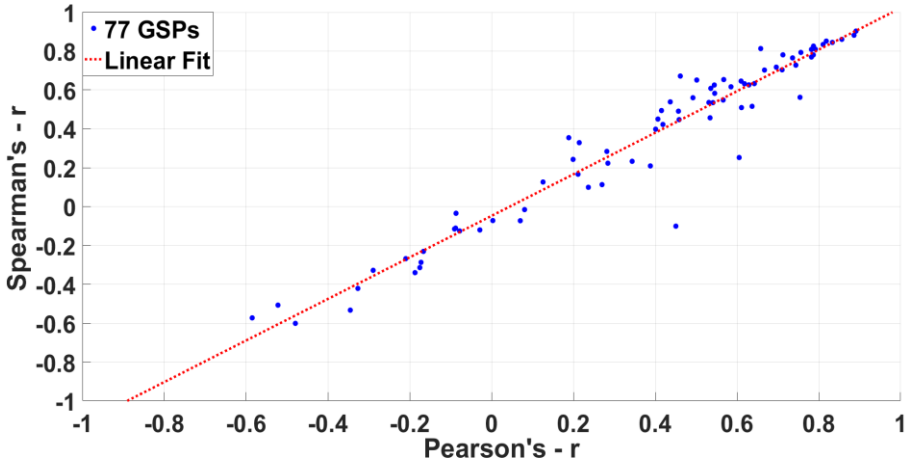


Figure 4.11: An example of correlation results based on the Pearson's and Spearman's coefficients, for active power with reactive power (average-daily values), for 77 GSPs

Figure 4.12 shows the results for active power (*P*) with: a) reactive power (*Q*), b) voltage (*V*), c) temperature (*T*), d) solar irradiance (*SI*), e) relative humidity (*RH*), f) atmospheric pressure (*AP*), g) wind speed (*WS*), h) solar azimuth angle (*A*) and i) solar elevation angle (*E*), based on the analysis of the average daily values, over the course of one year.

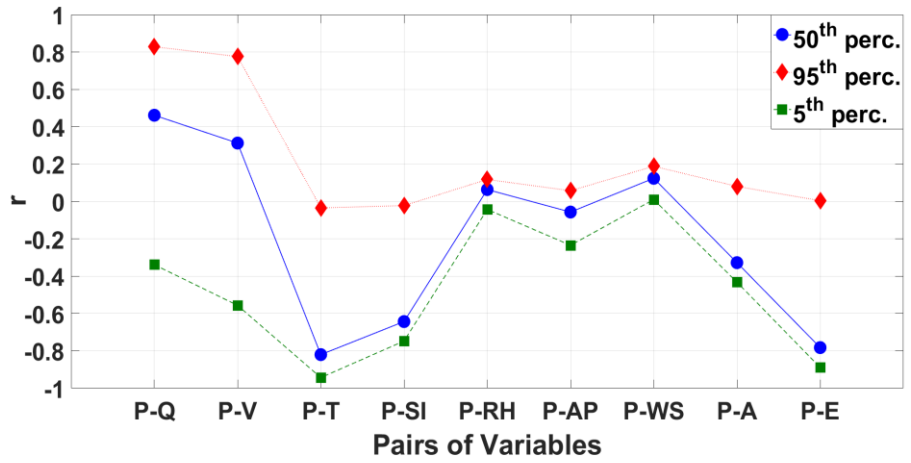


Figure 4.12: Seasonal correlations of average-daily values of active power (*P*) with: *Q, V, T, SI, RH, AP, WS, A, E*

No significant correlations are shown for active power with relative humidity, atmospheric pressure and wind speed. Active power is, in general, positively correlated with reactive power as indicated by the median and 95th percentile values, while the 5th percentile negative values show that the "expected" positive correlations are not present for a number of GSPs. This is also shown in more detail in Figure 4.10, where zero to negative correlations are demonstrated for, approximately, 25 % of the analysed sample. As shown in Chapter 3, Section 3.6, there is very good consistency in the seasonal profiles of active power, with decreasing demands for the summer period and increasing demands for the winter period. The same consistency in not present for reactive power (Figure 3.31) and this results in zero and negative seasonal correlations between *P* and *Q*. The correlations are stronger and far more consistent in the diurnal perspective, as shown in Section 4.4, Figure 4.6, which also shows that the strong diurnal correlations are relatively constant throughout the year. Furthermore, the seasonal correlations have a strong dependence on the selected period of the day for which the seasonal components are correlated, e.g. seasonal correlations at 20:00 hours are stronger than for 04:00 hours. This is demonstrated in more detail in the following section (i.e. 4.6).

For active power with voltage, in Figure 4.12, positive correlations are shown for the median and 95th percentile values (~0.3 and ~0.8), and negative correlations for the 5th percentile value (~-0.6). The extent of the resulting correlation coefficients indicates that similar inconsistencies are present, as with reactive power and that, based on the current analysis, no general characteristic associations can be established between the seasonal components (of average-daily values) of active power and voltage. It should be noted, that a more detailed inspection of the resulting seasonal correlations between *P* & *Q* and *P* & *V* has not revealed any potential groups according to customer-class percentages, i.e. the differences in these correlations are not suitable for GSP-classification, neither are these correlations indicative of particular customer sectors.

Regarding the meteorological and analemma parameters, active power has strong negative correlations with temperature and solar elevation angles (median values at -0.8) and to a lesser extent with solar irradiance and solar azimuth angles (median values -0.6 and -0.3, respectively). The small deviations between the median and the 5th percentiles indicate that, in contrast to *P* & *Q* and *P* & *V*, active power variations are consistently correlated with *T*, *SI*, *E* and *A*, for the majority of the GSPs. The fact that temperature and solar elevation angles are the best predictors of active power demand seasonality is further demonstrated in the per half-hour analysis (Section 4.6) and it is used in various instances in subsequent analysis.

Figure 4.13 (a) presents the seasonal correlations of average-daily values (average of 48 half-hourly measurements) of reactive power (*Q*) with: a) voltage (*V*), b) temperature (*T*), c) solar irradiance (*SI*), d) relative humidity (*RH*), e) atmospheric pressure (*AP*), f) wind speed (*WS*), g) solar azimuth angle (*A*) and h) solar elevation angle (*E*).

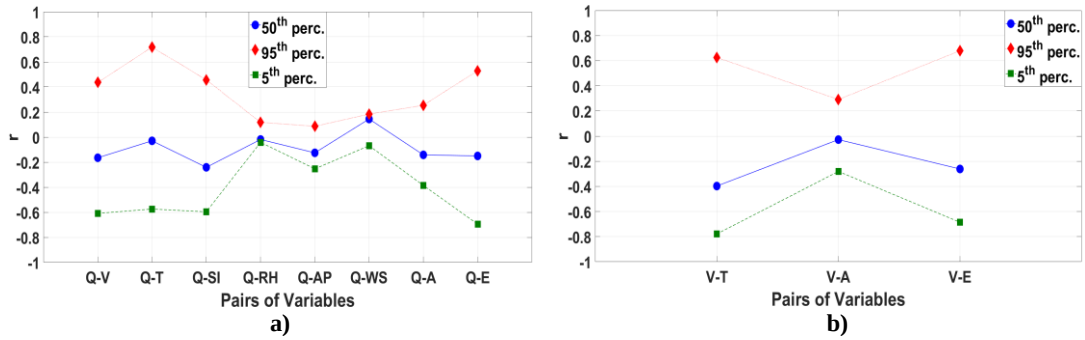


Figure 4.13: Seasonal correlations of average-daily values of a) reactive power (*Q*) with: *V, T, SI, RH, AP, WS, A, E* and of b) voltage (*V*) with: *T, A, E*

No significant correlations can be reported for reactive power and relative humidity, atmospheric pressure and wind speed for neither the 95th percentile maximum nor for the 5th percentile minimum values, while the median values are very close to zero, for all three pairs of variables. Correlation coefficients for reactive power and voltage are, as in the case of active power and voltage, spread from moderate positive values (~0.4) to moderate negative values (~-0.6), which demonstrates that there are no consistent relationships between the two, over all GSPs. Similar results are shown for reactive power and temperature, solar irradiance, azimuth and elevation angles and unlike active power, no correlations with reactive power have median values that largely deviate from zero, either positively or negatively (excluding reactive power with active power, which have been presented in Figure 4.12).

Figure 4.13 (b) shows the results for the correlations of average-daily values for voltage (*V*) with: a) temperature (*T*), b) solar azimuth angle (*A*) and c) solar elevation angle (*E*). Median values are at a level of ~-0.4 for temperature, ~0 for azimuth angle and ~-0.2 for solar elevation angle. The spread of the 95th and 5th percentile values and a more detailed inspection of the results has shown that, as in the case of reactive power, there is low consistency among the correlations from the various GSPs and that no strong dependencies can be established between the corresponding pairs of variables.

The current section quantified the level of seasonal correlations between *P*, *Q* and *V* and between these three parameters and the available meteorological and analemma variables, in terms of percentile values of the resulting coefficients, over all available GSPs. The results are therefore indicative of the presence and strength of associations between the pairs of variables studied. Strong and consistent correlations, in the seasonal cycle, can be reported between

active power and reactive power as well as between active power and temperature, solar irradiance and solar elevation/azimuth angles. Parameters *Q* and *V* do not have strong and well defined seasonal components of the same phase angle over all available GSPs, which was also demonstrated in Chapter 3 and as a result they correlate weakly with weather and analemma parameters. In contrast, seasonal active power levels appear to be primarily determined by weather conditions, either directly (e.g. heating loads) or indirectly (e.g. change in occupancy levels), a fact with is reflected in the strong correlations with temperature, solar irradiance and solar elevation angles. These conclusions are used in the following section, for the selection of pairs of variables as inputs in regression analysis, evaluated for the seasonal variations but on a per half-hour of the day basis.

4.6 Seasonal Correlations Per Diurnal Periods

Seasonal correlations can be evaluated on per half-hour of the day basis (or according to data resolution), in order to determine the level of associations between dependent and independent variables throughout the year, for particular diurnal periods i.e. half-hours of the day, thus increasing the resolution of the correlation analysis. The procedure can be summarised in the following steps:

- For all input variables, values are normalised with respect to the maximum of each dataset (3.3) and weekdays/weekends are separated (based on the results presented in Chapter 3). Dependent and independent variables are arranged into matrices with columns representing half-hours of the day and rows as the weekdays (1-261), or weekends (1-104) of the year.
- For each pair of dependent-independent variables, for each GSP (depending on data availability from Tables 3.1 and 4.1) and for each half-hour of the day the correlations are calculated using simple linear regression (least-squares algorithm discussed in Section 4.3). The resulting coefficients are: a) the coefficient of determination R^2 and b) the beta coefficient (gradient of the best fit line).
- For each pair of variables, the resulting coefficients from the analysis of all available GSPs are used to calculate mean, median, 95^{th} percentile and 5^{th} percentile values (for R^2 and beta), at each half-hour of the day.

Pairs of variables are selected based on the results from the seasonal correlations of average-daily values, presented in Section 4.5. Results are therefore presented for active power with:

reactive power, voltage, temperature, relative humidity[12], solar irradiance, solar azimuth angle and solar elevation angle, in Section 4.6.1 and for reactive power with: voltage, temperature, solar irradiance and solar elevation angle, in Section 4.6.2. The use of normalised values allows for comparable levels of the beta coefficients for all considered GSPs, as well as for an easier interpretation of the results. The sign of the beta coefficients corresponds to positive or negative correlations and the absolute normalised value is the per-unit change of the dependent variable with respect to the per-unit change of the independent variable.

4.6.1 Active Power Regression Analysis

Figure 4.14 shows the results for the linear regression analysis between active power and reactive power (R^2 and beta coefficients) for the seasonal correlations, on a per half-hour of the day basis and considering weekdays only.

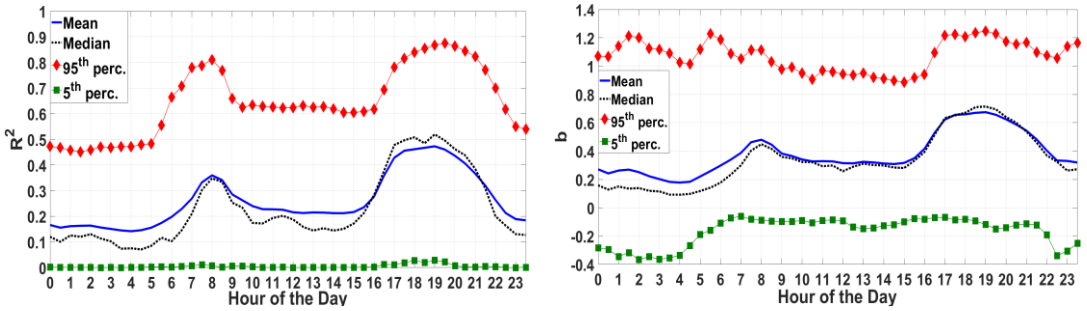


Figure 4.14: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with reactive power: a) R^2 and b) beta coefficients

Regarding mean and median values, the correlations are positive for the duration of the whole day and stronger during two specific periods, i.e. early morning (07:00 to 09:00 hours) and afternoon to evening (16:00 to 22:00 hours). As discussed in Chapter 3, Section 3.5 (Figure 3.22 (d)), these two periods are the ones with the highest levels of seasonal range-of-demand (normalised over the 48 half-hours of the day), i.e. periods of maximum seasonally variable demand with respect to the maximum demand at each half-hour. Another such period, for a number of GSPs, corresponds to the night hours (23:30 to 05:00), but weaker correlations are shown for active and reactive power during that period and thus it can be concluded that during the night the changes in active power demand cannot be (strongly) associated with changes in reactive power demand. Therefore, inferences about load-types can be made. For example, the decrease in correlations during the night indicates that the seasonally variable loads are

[12] Although relative humidity did not show strong seasonal correlations with active power in the previous section, it is often discussed in literature, primarily in load forecasting and it is therefore selected for further analysis in the current section.

constituted primarily of resistive components, e.g. thermal heating demands related to economy-7 meters, while the morning and evening periods (of strong correlations) must include loads that cause high reactive power variability such as lighting loads (CFLs), consumer electronics, wet loads, etc. These conclusions are further supported by the fact that the correlation profiles (R^2 values) are not the same for GSPs of characteristically different demand profiles, such as GSPs with primarily residential and commercial/mixture demands. A comparison between two such correlation profiles is presented in Figure 4.15, for GSPs 14 & 3. The contributions from the different customer-sectors, on which this discussion is based, have been estimated using the methodology presented in Chapter 5.

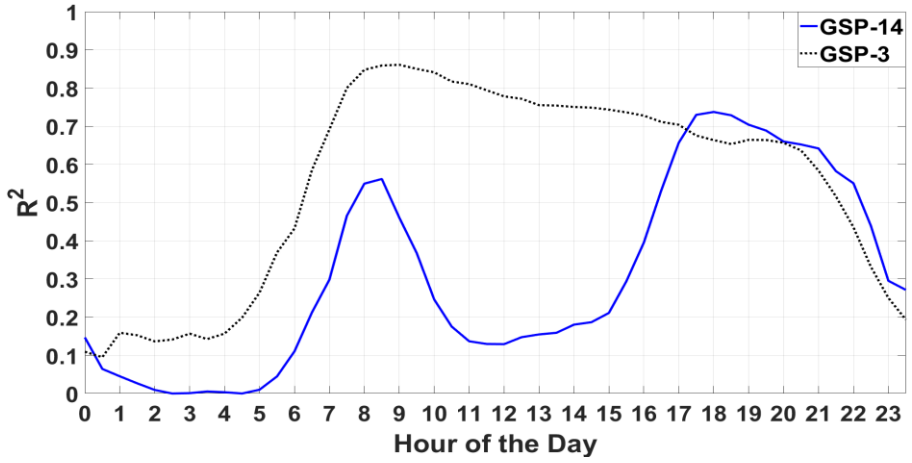


Figure 4.15: Correlation profiles (R^2 values) for GSPs with characteristic residential and commercial/mixture demand profiles

Weak correlations during the night as well as morning peaks are shown for both GSPs, but R^2 values are maintained relatively high for the commercial bus (GSP-3), while they drop during mid-day for the residential bus (GSP-14). The presence of an evening peak for GSP-3 indicates that, while this bus has high penetration of commercial demands during the period between 09:00 to 17:00 hours, it also has a significant portion of residential demand, as reflected in the strong correlations from 17:00 onwards and in the similar drop of the correlation coefficients, as GSP-14, for the period between 20:30 and 23:30 hours. The correlation profiles of individual GSPs can therefore be used for the assessment of load-types (disaggregation), for purposes of identifying portions of total demand from residential and commercial customers, as well as for classifying GSPs according to these characteristics and are therefore used in the following chapters of this thesis (and are not restricted to the dependencies between active and reactive power demands).

The 95th and 5th percentile values of the R^2 and beta coefficients, indicate that the correlation profiles may significantly vary for individual GSPs and can also be used to determine GSPs

of significantly different demand characteristics. Examples include GSPs with beta coefficients higher than 1, which means that for every 1 per-unit change in reactive power demand, more than 1 per-unit change is predicted for active power. This is not necessarily an indication of an outlier but such results need further investigation. The reasons for such discrepancies usually lie within the limitations of the ordinary least squares procedure. This is illustrated with an example in Figure 4.16, which corresponds to active/reactive power demands form a single GSP (GSP-19), throughout the year, at 17:00 hours.

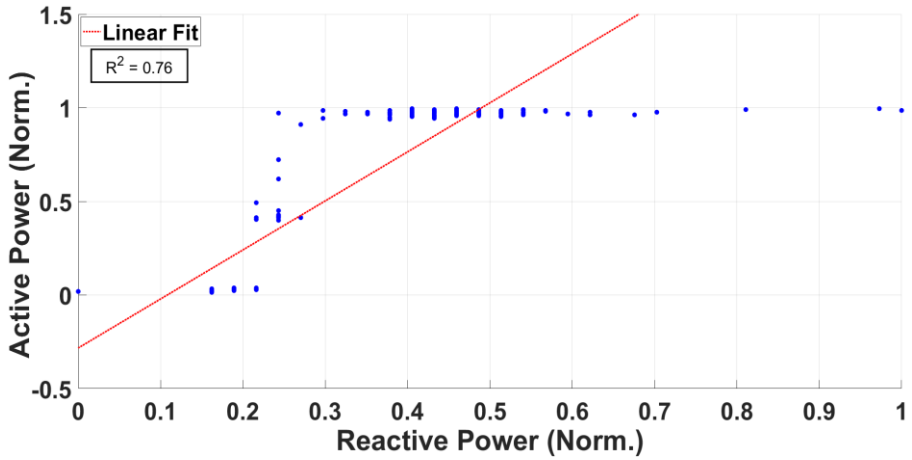


Figure 4.16: GSP-19 with ''atypical'' consumption characteristics and inappropriate use of the linear OLS fit

The linear regression gives a coefficient of determination of 0.76 and a beta-coefficient over 1. A visual inspection shows that a linear OLS fit does a poor job in estimating the underlying relationship. In such cases, higher degree polynomial fits are more appropriate, the disadvantage being that it is more difficult to interpret the (multiple) resulting beta coefficients (discussed in Section 4.3).

Figure 4.17 (a) shows the seasonal correlations, on a per half-hour of the day basis, between active power and voltage. No statistically significant correlations can be reported, with mean and median coefficients of determination below 0.1, for all half-hours of the day. Furthermore, the relatively close range of the 95th and 5th percentile values (from 0 to 0.3) indicates that the weak correlations are consistent among the 24 GSPs for which voltage measurements were available. These results are in contrast with the results presented in Chapter 3, Section 3.6, which demonstrated that voltage variability exhibits seasonal trends, at least for the majority of considered GSPs. The reason is that even though the day-to-day fluctuation of voltage, at particular half-hours, are within very narrow limits with respect to the nominal, these fluctuations are more pronounced than the seasonal component, as discussed in Chapter 3. As a result, the seasonality is "masked" and therefore correlations can be improved when

considering the filtered (smoothed) voltage values. These are shown in Figure 4.17 (b), in terms of the coefficient of determination and are based on smoothed voltage values calculated using of a moving-average filter of ± 2 weeks.

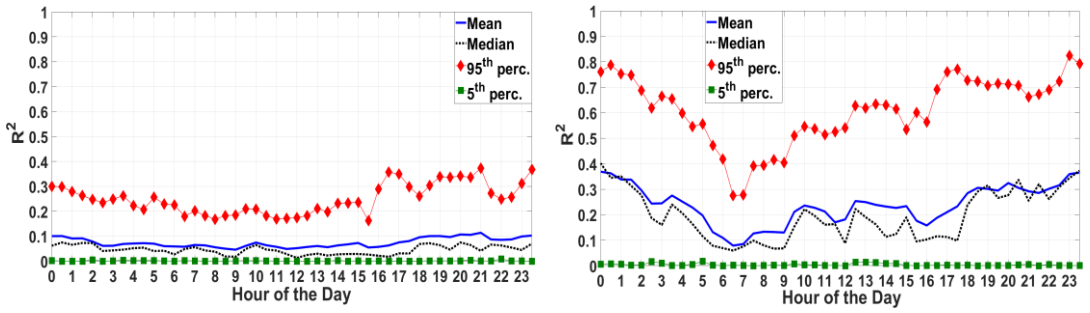


Figure 4.17: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with voltage – in R^2 values a) using actual values and b) using smoothed voltage values

The differences between the actual and smoothed values are demonstrated in the example shown in Figure 4.18, for GSP-58, at 17:00 hours. Despite the improved correlations in Figure 4.17 (b), these are still, generally, below $R^2 = 0.3$ and the results are highly inconsistent among the various GSP. Furthermore, and as it can be seen in Figure 4.18, this level of filtering produces smoothed values that largely deviate from the actual measurements and the results have very limited predictive value.

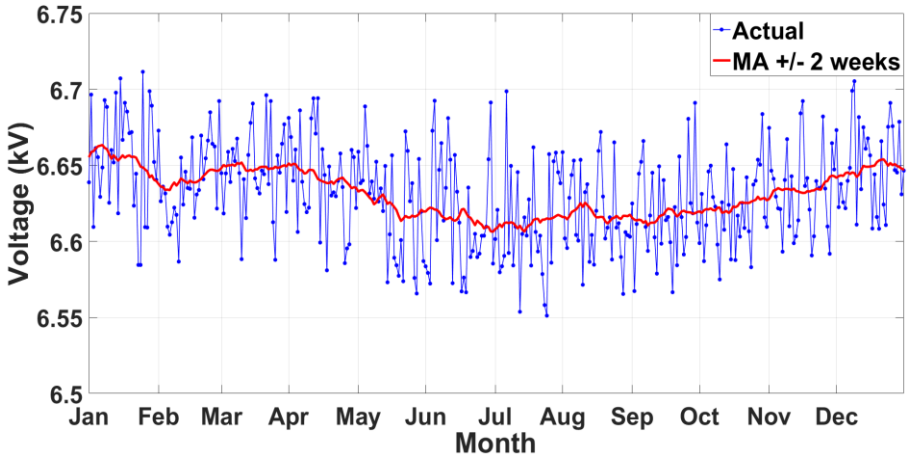
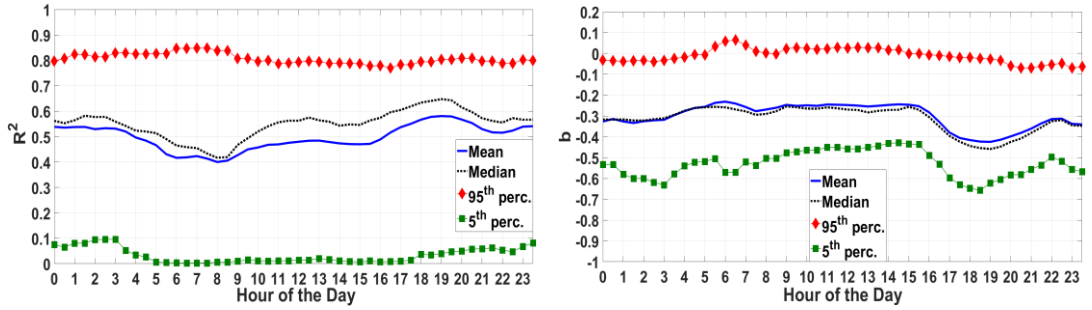


Figure 4.18: An example of moving-average smoothed values for voltage, at 17:00 hours, for GSP-58

Figure 4.19 presents the results for the seasonal correlations of active power and temperature, on a per half-hour of the day basis and considering weekdays only. Active power demand and temperature are negatively correlated with mean and median beta values between ~-0.2 to ~-0.5 and R^2 values at an average level of ~0.5 to ~0.6. The extended range of 95th and 5th percentile values is due to the fact that temperature dependencies are stronger for GSPs with

certain characteristics, such as mostly/primarily residential demands (further discussed in Chapters 5 and 6) and can reach levels of up to 0.8 and 0.9 (R^2) for several GSPs, without the use of filtering, even though correlations can be significantly improved by using the smoothed seasonal components of the two variables.



a) b)

Figure 4.19: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with temperature: a) R^2 and b) beta coefficients

The rate of change of active power demand with respect to the changes in temperature is increasing for the period between 16:00 to 22:00 hours, i.e. for the duration of time when people are at home and can be more responsive to day-to-day temperature changes. It is assumed that the correlations between the two variables includes components not only related to demand for thermal energy, i.e. temperature levels potentially affect occupancy levels and thus the use of home equipment, consumer electronics etc., due to the fact that during cold periods people tend to stay home for more prolonged periods of time. The drop in the correlations between 06:00 to 10:00 hours is assumed to be a result of the, relatively, fixed schedule of customers during the early morning period and their associated energy use, which is not as sensitive to temperature levels to the extent that it is during other periods of the day.

Figure 4.20 shows the differences in R^2 values between all days, weekdays and weekends. The period of maximum differences in active power demand between weekdays and weekends is concentrated around the early morning hours, but also extends up to 19:00-20:00 hours (shown in Chapter 3, Figure 3.27). However, temperature levels do not vary significantly between weekdays and weekends, which results in a disassociation of the two variables during these periods. The use of both datasets (denoted as All Days) compromises the linear-fit by increasing the residuals, due to the inconsistencies between the weekday-weekend demand levels.

It should also be noted that the geographical locations from which demand and weather data were obtained have similar climate characteristics corresponding to temperate oceanic and temperate continental climates, associated with longer and colder winters and temperate summers. As a result, the effects of summer air-conditioning loads are significantly lower for

what can be expected for locations experiencing warmer summers. For the analysis of demands from such locations the regression approach needs to be modified to account for the increasing demands for AC loads.

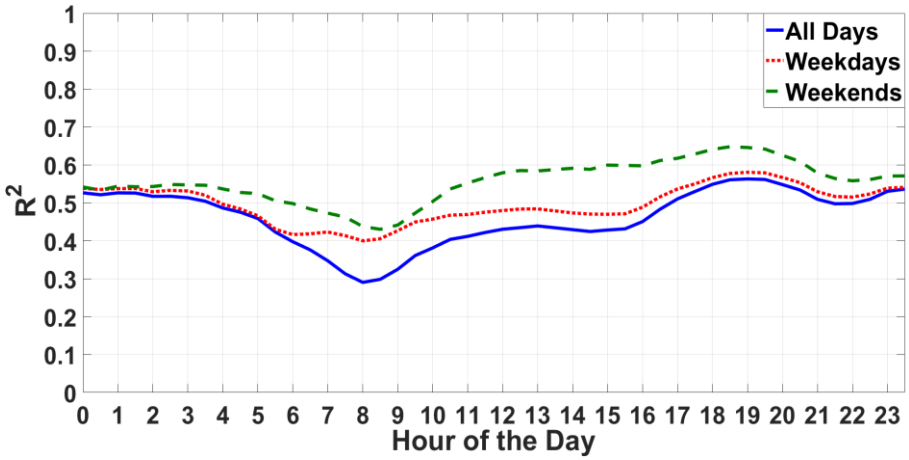


Figure 4.20: Coefficient of determination - R^2 per half-hour of the day for active power with temperature and for all days, weekdays and weekends

Figure 4.21 show the results for the seasonal, per half-hour correlations between active power demand and relative humidity (weekdays only). The coefficient of determination is below 0.2 for all half-hours of the day, in (a). However, by applying a moving-average filter (Section 4.3) of ± 2 weeks to both datasets, the correlations can be significantly improved reaching values of up to 0.5 for particular periods of the day, i.e. between 02:00 to 08:00 hours and between 16:00 to 20:00 hours, as shown in (c). The beta coefficients, for the non-filtered data in (b), alternate between negative and positive with the highest values during the same two periods.

Higher percentages of relative humidity in the atmosphere indicate higher levels of partial pressure exerted by water vapor and can therefore reduce the rate of the evaporation process. One of the mechanisms for body temperature regulation, in humans, is perspiration and the subsequent evaporation of sweat from the skin and therefore relative humidity affects the way temperatures are perceived, i.e. higher relative humidity causes more apparent heat at constant temperatures. However, the effects of relative humidity alone become more complicated because it is itself a function of temperature, i.e. saturation levels increase for increasing temperatures and thus relative humidity drops, provided no more moisture is added to the air. Metrics such as the "heat index" are designed to account for the combined effects of air temperatures and humidity levels, discussed in [209], [210] and [211]. For the results presented in Figure 4.21 (b) the interpretation is as follows: During night hours increase in relative humidity is associated with an increase in temperature levels (based on the available data and

the regression analysis presented here), so the two weather variables are positively correlated during these periods, as it can be seen from the beta coefficients for temperature and relative humidity in Figure 4.21 (d). Because temperature is negatively correlated with active power demand (on the seasonal perspective), the same relationship is evident for relative humidity as well. However, during the day, temperature and relative humidity appear to be negatively correlated (negative beta coefficients in Figure 4.21 (d)) and thus active power and relative humidity are positively correlated.

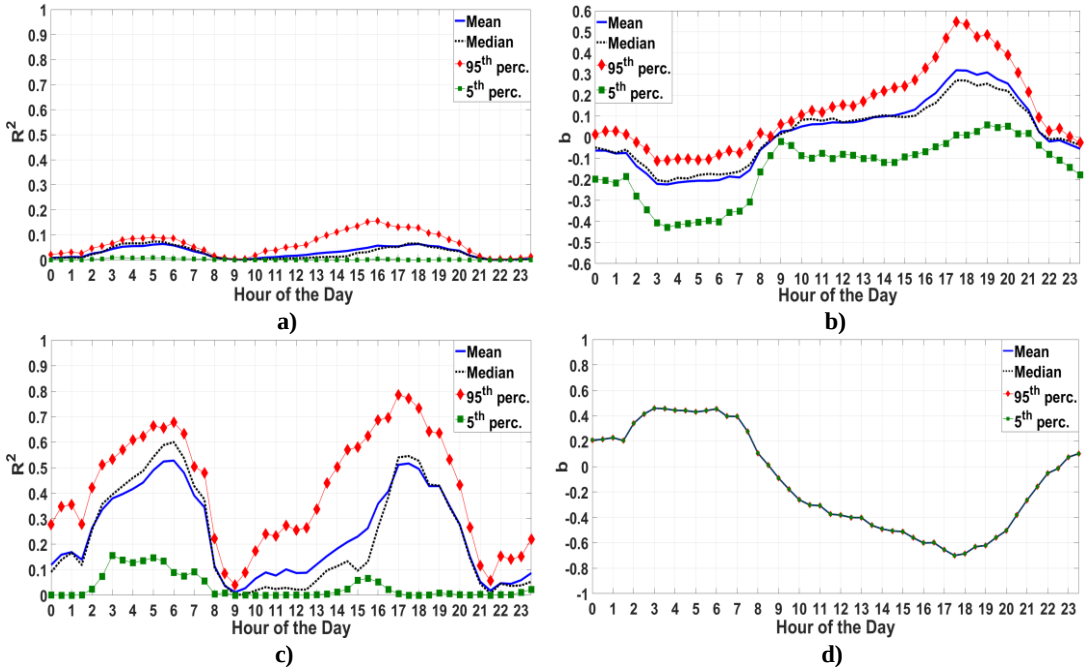


Figure 4.21: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with relative humidity: a) R^2, b) beta coefficients, c) R^2 for filtered data and d) beta coefficients for temperature and relative humidity

Whether these results can be exclusively attributed to issues of collinearity between temperature and relative humidity needs further investigation to be established but, in any case, the correlation coefficients for actual measurements are not statistically significant to justify further analysis. The effects of relative humidity on perceived temperatures and subsequently on active power demand, are expected to increase when considering datasets from different geographical locations (e.g. from warm-humid climates).

In Figure 4.22, the results for active power with solar irradiance are presented. The values are limited to the periods of the day when solar irradiance measurements are available (during sunlight hours) and the strength of associations is generally moderate to low, with correlation coefficients reaching maximum values of ~0.3 (mean/median) and ~0.5 (95^{th} percentile) at around 07:30 and 19:00 hours. The correlations are negative, indicated by the values of the beta coefficients, which are within 0 and ~-0.4 for most periods of the day, but rabidly

increasing for dawn and dusk hours. These are the periods of shifting sunlight levels throughout the year and it is assumed that they directly affect demand for lighting loads, while also having effects on occupancy related loads. The hypothesis is also supported by the results presented in Chapter 3, Section 3.8, where it has been demonstrated that seasonal-diurnal profiles of the rate of change of active/reactive power demands have characteristic peaks of increasing demand levels from 16:00 to 22:30 hours, which mark the yearly solar irradiance levels during the same periods and reach maximum and minimum excursions at the times of the summer and winter solstices.

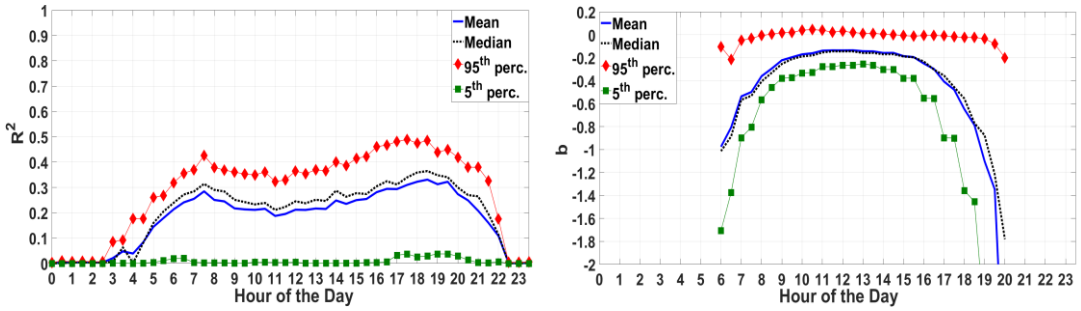


Figure 4.22: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with solar irradiance: a) R^2 and b) beta coefficients

In Figure 4.22 (b) the beta coefficients are presented from 06:00 hours and up to 20:00 hours because the limited number of non-zero data-points for half-hours of the day outside these periods reduces the validity of the regression analysis and produces unrealistically high beta coefficients (with corresponding R^2 values of ~0).

The problems regarding zero or non-available solar irradiance measurements can be resolved by the use of the solar elevation angles, which mark the position of the sun in the sky (altitude), for the selected periods of the day, throughout the year (Section 4.2). The regression analysis results for active power demand with solar elevation angles are presented in Figure 4.23.

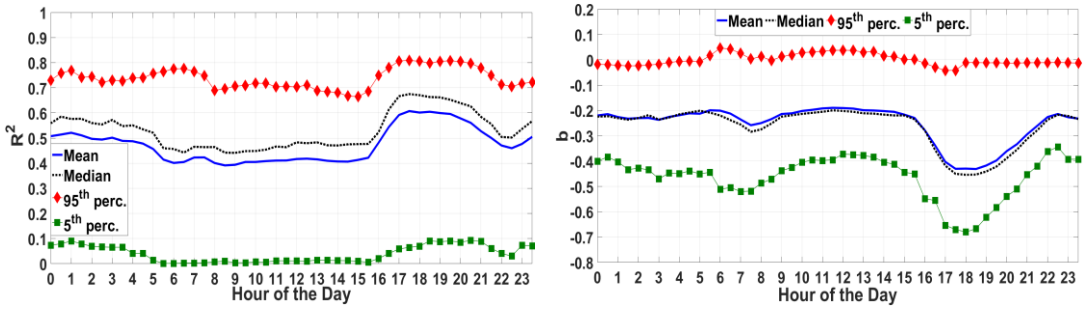


Figure 4.23: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with solar elevation angles: a) R^2 and b) beta coefficients

Because solar elevation data is unaffected by atmospheric phenomena, the variables capture seasonality as a "smooth" continuous variable and are therefore highly correlated with active power demand, comparable (or higher) to the levels of correlation with temperature measurements. R^2 values (mean/median) range from ~0.4 to ~0.7, with a peak between 16:00 to 22:00 hours, which is more pronounced than the corresponding peak for active power demand with temperature, while 95th percentile values have the same diurnal pattern and reach values of ~0.8. Beta coefficients are negative, indicating decreasing active power demand for increasing solar elevation angles, with a peak during the same period of the day as the coefficient of determination, as well as with the results presented for the temperature correlations (Figure 4.19). The beta values (mean/median) also reflect the results for solar irradiance, Figure 4.22 (b), for the mid-day period for which continuous solar irradiance measurements are available.

Since solar elevation angles are good indicators of seasonality, the evening peak observed for R^2 and (negative) beta-values indicates that this is the period of maximum seasonally variable demand, while the early morning variability in active power (Chapter 3, Figure 3.22) is mostly associated with the changes in demand between different days of the week and the variations are to a lesser extent the result of seasonal changes. This is also reflected in the rate of change of active power (Chapter 3, Figure 3.42) which is at constant levels and does not shift throughout the year, for the corresponding morning period. Conversely, the rate of change for the evening hours is shifting according to seasonal changes, as mentioned in the discussion for solar irradiance correlations.

Figure 4.24 shows the linear regression results for the per half-hour seasonal correlations of active power and solar azimuth angle. As discussed in Section 4.2, solar azimuth angles mark the relative position of the Sun in the sky with respect to the east-west axis.

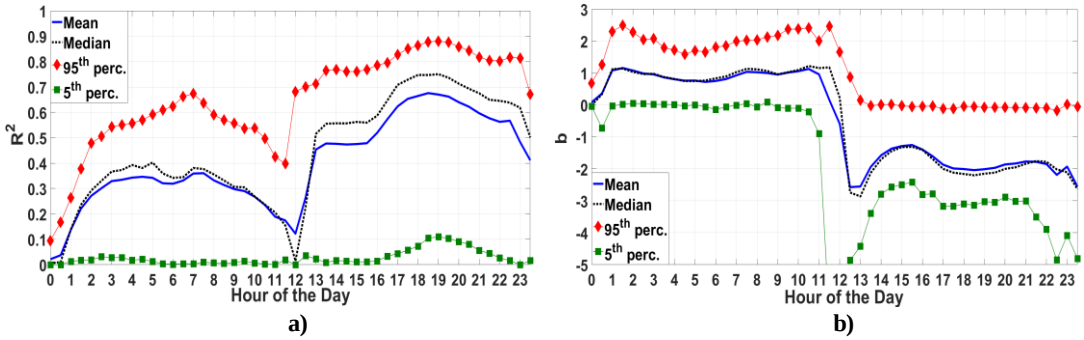


Figure 4.24: Linear regression results - seasonal correlations on a per half-hour of the day basis for active power with solar azimuth angles: a) R^2 and b) beta coefficients

The decrease in the strength of correlations at the mid-day period can be explained with reference to Section 4.2. Solar azimuth angles at selected hours of the day can be visualised as

the x-axis component of the individual "8-shapes" shown in Figure 4.2. During mid-day, the inclination of these "8-shapes" is at its minimum and therefore the corresponding variable's ability to mark seasonality is substantially compromised, i.e. there is minimum deviation (in degrees) between the extreme points (summer and winter solstices). This is also the reason for the shift of beta-coefficients from positive to negative, which occurs at around 12:00 hours. For the remaining periods of the day, correlations are between 0 and ~0.7, reaching a peak during the afternoon/evening hours, as in the case of temperature, solar irradiance and solar elevation. Because of the inconsistencies mentioned above, in subsequent analysis, the use of solar analemma variables is restricted to the solar elevation angles, apart from the multiple regression forecasting model, presented in Chapter 7.

4.6.2 Reactive Power Regression Analysis

Figure 4.25 shows the results for the per half-hour seasonal correlations between reactive power and voltage. As in the case of active power and voltage, correlations are very weak (generally below ~0.3) and the reasons are same as outlined in the discussion provided in Section 4.6.1, Figures 4.17 and 4.18.

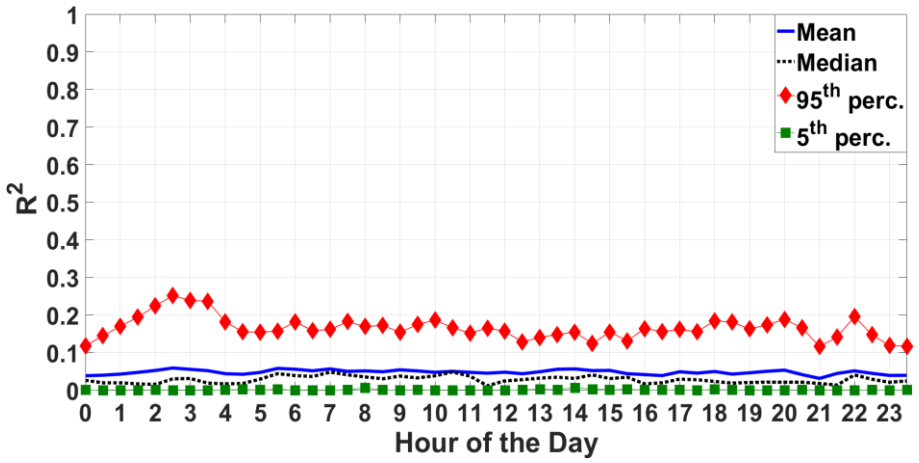
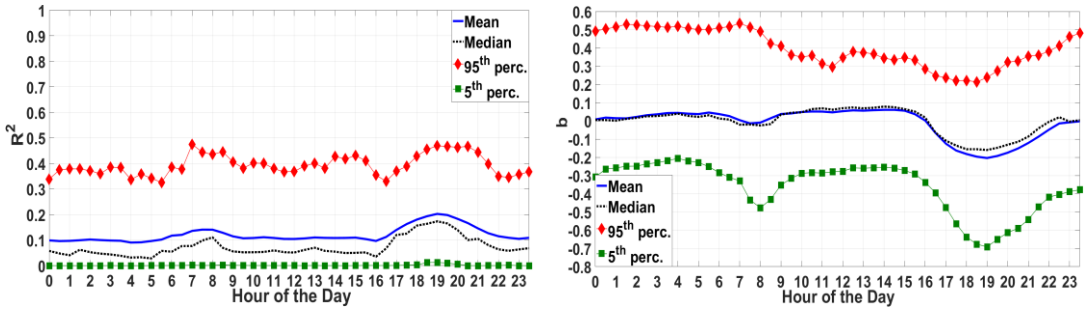


Figure 4.25: Linear regression results - seasonal correlations on a per half-hour of the day basis for reactive power with voltage – R^2

In Figure 4.26 the per half-hour seasonal correlations between reactive power and temperature are presented. R^2 values are below ~0.3 for the mean and 50th percentile, however and as shown from the 95th percentile values, for a number of GSPs the coefficient of determination is up to ~0.5. Beta coefficients are negative apart from the 95th percentile and the characteristic peaks during morning periods (07:00 to 09:00 hours) and afternoon/evening periods (16:00 to 22:00 hours) are lower than the corresponding peaks shown for the active power-temperature analysis. Issues of multicollinearity (as discussed in Section 4.3) apply here as well, since the

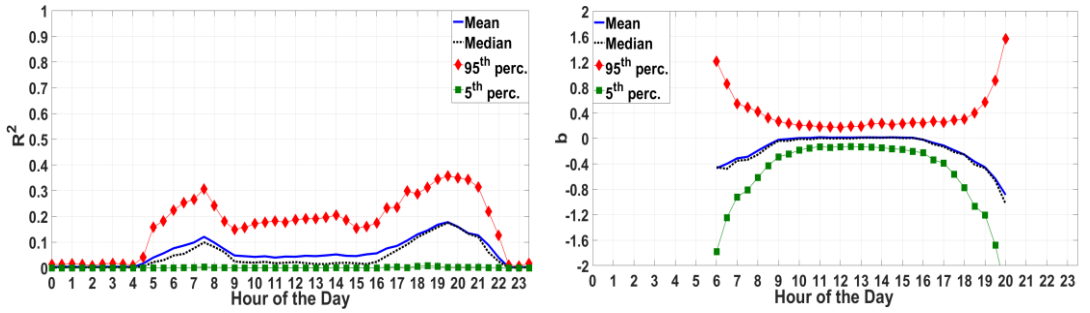
moderate/strong positive correlations between active power and reactive power (Figure 4.14) and the strong negative correlations between active power and temperature (Figure 4.19) would have inevitably resulted in correlations between reactive power and temperature (based on simple linear regression analysis). Nevertheless, the differences in the strength of these correlations indicates that weather conditions primarily affect active power and the effects on loads requiring both active and reactive power are weaker. This is also shown based on the results for reactive power with solar irradiance and azimuth/elevation angles, in the following figures.



a) b)

Figure 4.26: Linear regression results - seasonal correlations on a per half-hour of the day basis for reactive power with temperature: a) R^2 and b) beta coefficients

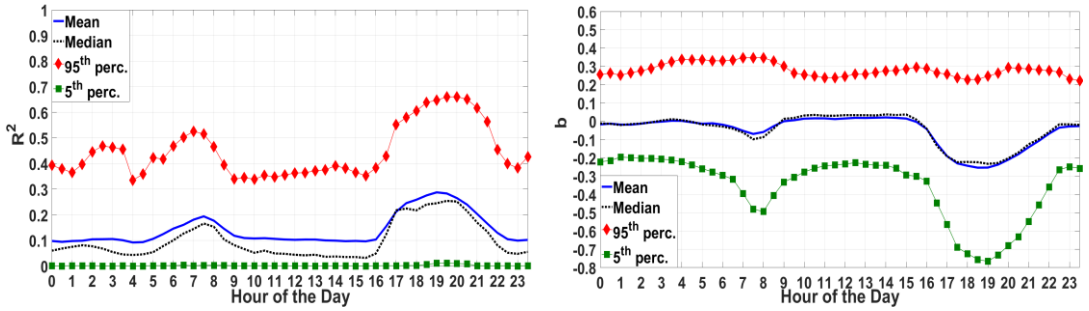
Figure 4.27 shows the results for reactive power and solar irradiance. Only weak correlations can be reported for the mean/median values, while the 95th percentile values indicate that, for a number of GSPs, the correlations reach values of ~0.4 R^2. The profiles for the beta-coefficients are similar to the ones shown for active power and solar irradiance in shape, but with zero values throughout most of the day and with peaks for the periods during early morning and evening hours. The increase of the beta coefficients for the evening hours is in agreement with the discussion presented in Chapter 3, Section 3.8, for the rate of change of reactive power and the seasonality pattern which extends from 16:00 hours during winter, to 22:00 hours during summer (similar to the correlation patterns shown for active power and solar irradiance).



a) b)

Figure 4.27: Linear regression results - seasonal correlations on a per half-hour of the day basis for reactive power with solar irradiance: a) R^2 and b) beta coefficients

The results for the seasonal correlations of reactive power demand and solar elevation angles are shown in Figure 4.28. Correlations are improved compared with the results obtained using solar irradiance, for the reasons mentioned in the discussion for active power demand and solar elevation angles, in Figure 4.23. Mean and median beta-coefficients are at zero level and increase (negatively) only for the morning and evening hours. The coefficient of determination, however, for mean/median values is below statistically significant limits (generally below 0.3), but as it can be seen from the 95th percentile values it is between ~0.4 and ~0.7, for a number of GSPs.



a) b)

Figure 4.28: Linear regression results - seasonal correlations on a per half-hour of the day basis for reactive power with solar elevation angles: a) R^2 and b) beta coefficients

In Sections 4.6.1 and 4.6.2, the seasonal correlations between active power and reactive power, as well as between these parameters and available meteorological and analemma variables have been presented, on a per half-hour of the day basis, that allows to investigate the variability in the level of seasonal correlations within the day. For active power, significant correlations can be shown with temperature, with the analemma variables and to a lesser extent with solar irradiance levels. These correlations have diurnal profiles that reflect the changing levels in the sensitivity of customers' responses to weather condition and open the possibility for load disaggregation based on a per half-hour of the day analysis.

Reactive power shows weaker correlations with the above mentioned external conditions; the diurnal correlation profiles do, however, show patterns and maximised correlations for specific periods of the day, which can be associated with seasonal changes in specific load-categories that affect the requirements for reactive power. In the same context, the correlation analysis between active and reactive power opens similar possibilities for load disaggregation, as the resulting diurnal patterns indicate periods of common changes in *P*-*Q* levels and periods of non-correlated changes. These distinctions are therefore further discussed in Chapter 6.

The strength of active power correlations with weather variables, or other markers of seasonality, such as solar elevation angles, can be shown to be associated with the GSP

composition according to customer-class percentages[13], as shown in Figure 4.29 (a). Similar associations are not possible based on reactive power correlations, as shown in Figure 2.29 (b). This shows that the seasonal components of reactive power (discussed in Chapter 3), are not consistently in phase with markers of seasonality.

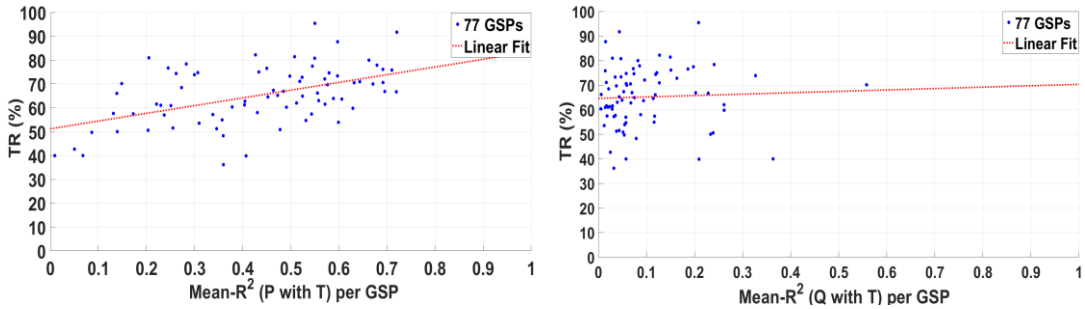


Figure 4.29: Average R^2 values of linear regression between a) active power and temperature and b) reactive power and temperature, with respect to percentages of total residential demand

Regarding voltage, no significant correlations can be shown with neither active power, nor reactive power (nor with any of the meteorological and analemma parameters). The reason is the significant stochastic components in voltage variability and a possible solution is the use of smoothed values. This, however, results in correlations that are not truly representative of the day-to-day covariance of voltage with other parameters.

4.7 Analysis of Residuals and Multiple Regression

The assessment of the "goodness-of-fit", for the individual linear regression models presented for the per half-hour analysis in Section 4.6, is quantified by the coefficient of determination - R^2, which is a function of the residual sum of squares (RSS). Analysis of the residuals themselves can be used to account for problems of heteroscedasticity and non-linearity and reveal the particular weaknesses of the linear regression models, which can then be used to determine different approaches to improve modelling performance, or to interpret the results in ways that are useful in understanding the underlying relationships between electricity demand and external conditions.

Figure 4.30 shows the distribution of residuals for the linear regression analysis between active power and temperature. The individual lines in the plot represent the 48 half-hours of the day

[13] These are estimated based on the approach presented in Chapter 5. While 98 GPSs are available for *P* with *T*, only 77 GSPs are available for *Q* with *T* and therefore, for easier comparisons between the two plots, only 77 GSPs are presented.

and correspond to the mean residual values over all analysed GSPs, for each weekday of the year.

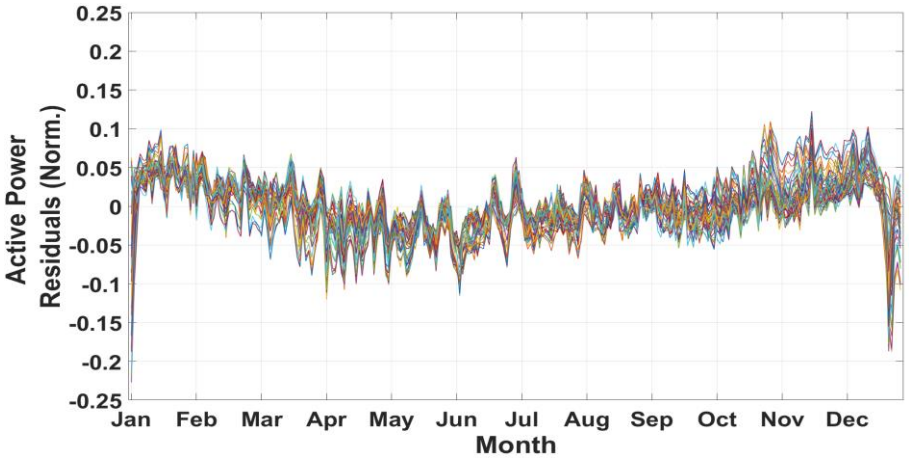


Figure 4.30: Seasonal distribution of residuals: linear regression analysis of active power with temperature, averaged over all GSPs, weekdays only

A general trend of positive residuals for the winter months and negative residuals for the summer months can be observed. The large deviations from zero at the edges of the plot correspond to the Christmas holiday season (particularly at 25th December and 1st January), an indication that model performances can be improved by the removal of specific "special days" from the analysis. The input datasets are normalised with respect to the maximum values (3.3) and thus the residuals are presented in normalised values as well.

A specific example of the residuals of a linear regression model (active power with temperature) for GSP-47, at 17:00 hours is presented over a three-year period in Figure 4.31 (a). Figure 4.31 (b) shows the sample autocorrelation, which can be used to detect periodic signals in the sample, that can be, potentially, obscured due to stochastic components (also mentioned in Chapter 3, Section 3.3 as an alternative to Fourier analysis for detecting periodicities in the available MV measurements).

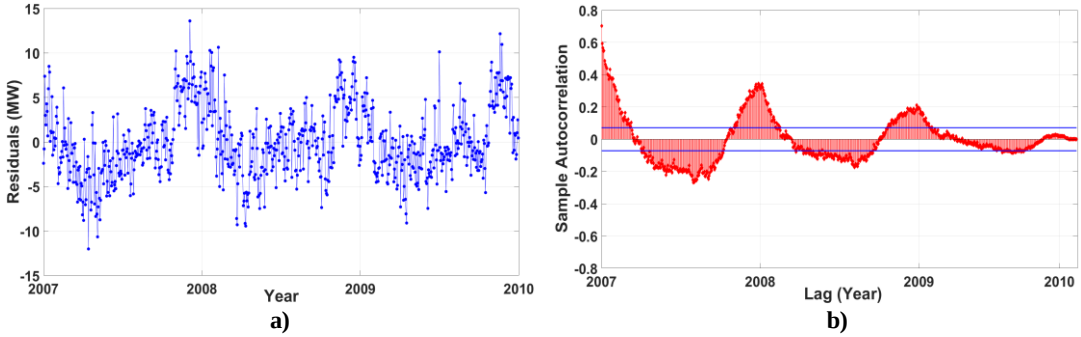


Figure 4.31: Residuals of linear regression analysis of active power with temperature for 3-years at 17:00 hours: a) actual residuals (MW) and b) sample autocorrelation, (weekdays only)

The seasonality of the residuals has a period of approximately one calendar year with positive and negative peaks during the mid-winter and mid-summer periods. However, the maximum deviations from zero are slightly shifted with respect to the peaks of the autocorrelation function, i.e. higher negative residuals around March-April and higher positive residual around October-November.

This is better illustrated by the use of an active power and temperature scatter-plot of weekdays only, at the same daily period (i.e. 17:00 hours) and using the moving-average filtered values of ± 2-weeks, as shown in Figure 4.32 (a). Marked on Figure 4.32 (a) is an example of the differences in active power demands at periods of relatively similar temperature levels. In Figure 4.32 (b), the differences in solar elevation angles (reversed axis) are presented with respect to periods of same solar azimuth angles. Apart from the apparent similarities in the "8-figure" shapes of the two plots, a useful observation is that the periods for which there are higher deviations between expected (from a linear fit) and actual active power demands, for same temperature levels, coincide with the periods of maximum deviation of solar elevation angles.

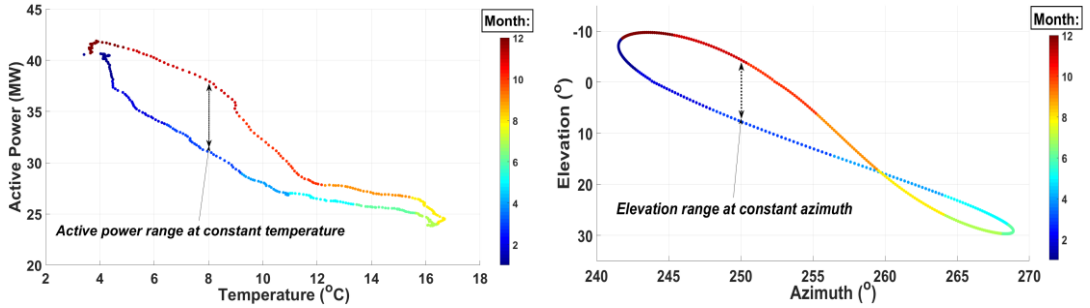


Figure 4.32: Moving-average filtered values of yearly: a) active power demand and temperature and b) solar elevation and azimuth angles, at 17:00 hours

The characteristic "loops" shown for both relationships are, from a mathematical perspective, a result of the phase difference of the yearly time-series of the presented variables. This is shown in Figure 4.33 using the z-score (3.1) normalised values of active power, temperature, solar elevation angle and solar azimuth angle, calculated, in the case of active power and temperature, for the smoothed yearly values. For active power, the axis is reversed to allow for easier comparison with the other variables.

The two periods marked on Figure 4.33, i.e. black-dashed lines, correspond to similar temperature levels (-0.5 in z-scores) and are located at mid-March and mid-November. At these periods, active power demand is at levels of 0 and -1.12 respectively. There is, therefore, higher demand in mid-November than in mid-March, even though temperature is (on average) the same. The corresponding solar elevation levels (which would exactly match solar

irradiance if it was not for atmospheric conditions) are higher for mid-March than for mid-November and the differences with temperature are positive, in the first instance, and negative in the second. These results suggest that during periods of matching temperature levels, active power variability can be further explained with respect to solar elevation angles, as well as with respect to other explanatory variables.

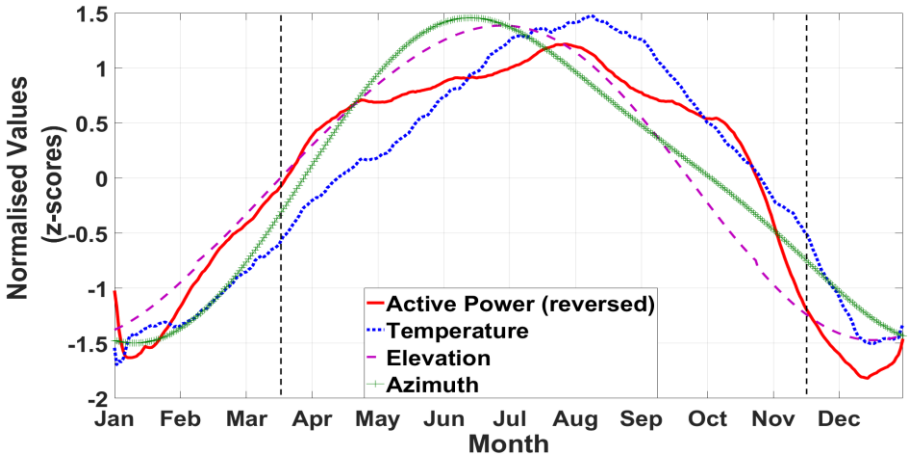


Figure 4.33: Seasonal components of 3-year averaged values of active power and temperature and seasonal components of solar elevation/azimuth angles

This hypothesis can be further supported by the resulting correlations of the residuals of active power demand and temperature, with reactive power, solar irradiance and solar elevation angles, as presented in Figure 4.34.

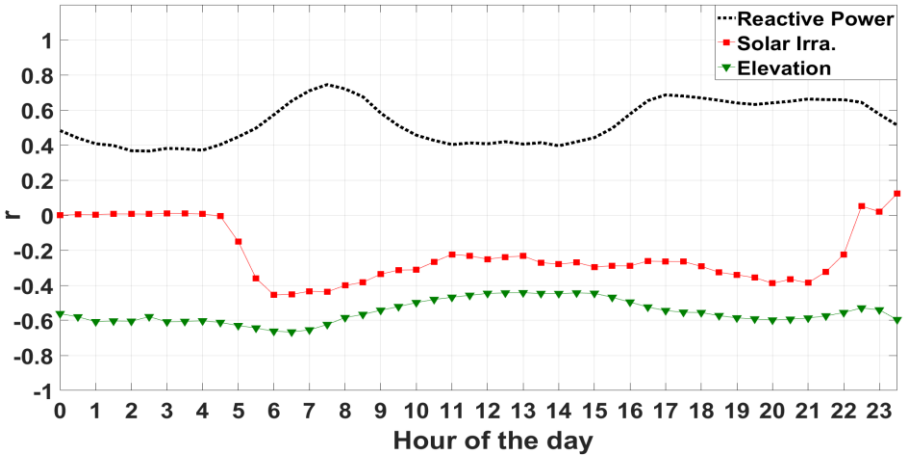


Figure 4.34: Correlations of the residuals of linear regression between active power and temperature with: reactive power, solar irradiance and solar elevation angles

The residuals from the linear regression of active power demand and temperature (3-years of data at all half-hours of the day) are strongly and positively correlated with reactive power, strongly and negatively correlated with solar elevation angles and, in the case of solar

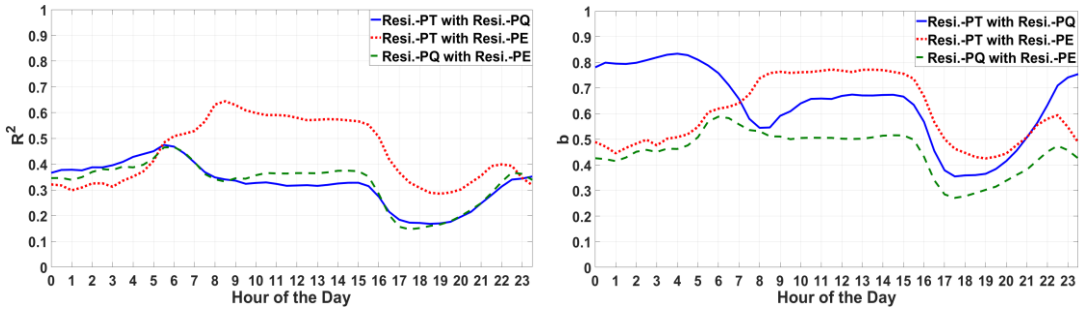
irradiance, the residuals are negatively correlated with the resulting diurnal patterns matching the results for solar elevation angles, but with lower correlation coefficients (for the half-hours of the day for which solar irradiance measurements are available). These correlations suggest that active power demand differences at constant temperature levels are accompanied by changes in reactive power and by the opposite changes in solar elevation angles, i.e. underestimations of active power demand values based on active power-temperature linear fits correspond to periods of higher reactive power levels and vice-verse, while these underestimations correspond to periods of reduced solar elevation angles (and thus solar irradiance levels) and vice versa. Inferences about the types of loads responsible for these changes can therefore be made, based on the fact that they correlate well with solar irradiance levels and reactive power and may include lighting loads, consumer electronics, wet loads, etc.

The above analysis is limited in the sense that the correlations between the presented variables, in the seasonal time-frame, have already been established (Sections 4.5 and 4.6) and are therefore expected, even in the absence of particular load types that can support them. For example, because active and reactive power demands are strongly and positively correlated, the residuals of active power with temperature would have been strongly correlated with reactive power even without any specific load-type associations. Furthermore, while for the referenced periods, i.e. mid-March and mid-November, the temperatures are at approximately the same levels, the rate of change of temperature is positive and negative, respectively. Therefore, the different levels of active power demands may be, at least partly, attributed to psychological effects that determine demand for different load-types, i.e. solar irradiance levels may affect the subjective perception of ambient temperature and also the "path" from summer-to-winter and winter-to-summer may affect the overall demand for heating/cooling loads, due to the change in temperatures (during those period) and not due to the actual temperature levels.

Another way to examine these relationships and demonstrate that the changes in active power demands at constant temperature levels are, potentially, due to particular load types, is to perform correlation analysis of the residuals of three different linear regression models and based on all available GSPs, on a per half-hour of the day basis, i.e.:

1. Residuals of active power and temperature with the residuals of active power and reactive power – (*Res.-PT* vs *Res.-PQ*)
2. Residuals of active power and temperature with the residuals of active power and solar elevation angles – (*Res.-PT* vs *Res.-PE*)

3. Residuals of active power and reactive power with the residuals of active power and solar elevation angles – (*Res.-PQ* vs *Res.-PE*)



a) **b)**

Figure 4.35: Correlations of the residuals of various linear regression models (for active power analysis)

The corresponding results are presented in Figure 4.35 and can be interpreted in the following way: the residuals of the simple linear regression models essentially represent the deviations of active power from what the least squares method has established the best description of the linear relationship to be (in terms of minimizing the sum of the squared errors). Therefore, for each explanatory variable (temperature, reactive power and elevation angle), the residuals correspond to the points which fail to be predicted, when assuming that the underlying relationship is linear. If the residuals of two such models correlate strongly, it is an indication of a commonality of errors and collinearity between the independent variables. This is related to approaches such as partial-correlation analysis, as applied to linear regression (discussed in Section 4.3). Consider *Res.-PT* vs *Res.-PQ* (blue solid line in Figure 4.35). Active power data-points above what is expected for a given temperature level are also above what is expected at a given reactive power level (when the correlations of these residuals are significantly high and positive). When the residuals of two regression models are not strongly correlated, it is an indication of reduced multicollinearity and vice-versa.

For example, Figure 4.36 shows the correlations of *Res.-PT* vs *Res.-PQ*, for two GSPs (GSPs-14 and 3 as presented in Section 4.6, Figure 4.15) which correspond to a predominantly-residential and a commercial/mixture GSP, respectively. The correlations of the residuals are higher for the commercial GSP while there are generally low for the residential GSP, with increased correlations only during night hours. This indicates that the portion of variance of active power not explained by both independent variables is shared (or common) to a higher degree for GSP-3 than for GSP-14. These portions drop (for both GSPs) for the periods of the day when demand levels are mostly determined by people's daily schedules (early morning period) and when the errors are not commonly shared (or correlated), such as during the evening period (between 17:00 and 21:00 hours). These distinctions are important not only

from a theoretical perspective, e.g. regarding the evaluation of multicollinearity effects, but also from a practical point of view, because they need to be taken into account when developing disaggregation approaches. The results are therefore further discussed in Chapter 6 regarding the selected approaches for load-type identification.

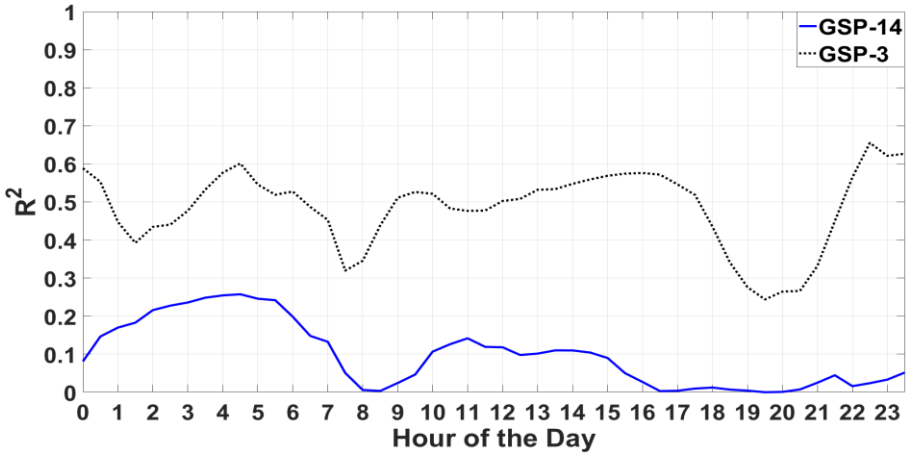


Figure 4.36: Example of the correlations between the residuals of active power with temperature and active power with reactive power, for two characteristic GSPs

Regarding the periodicities of the residuals, as presented in Figure 4.31, it should be noted that these are not a result of restricting the linear regression models to a 1st degree polynomial best fit. These seasonal patterns are (generally) evident even when using higher degree polynomials, such as of the 2nd or 3rd degree, as shown in Figure 4.37.

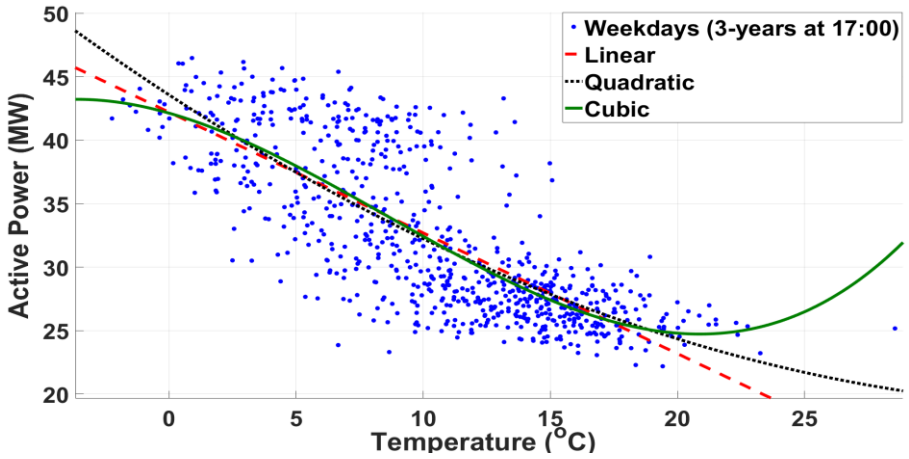


Figure 4.37: Comparisons of linear regression of the 1st, 2nd and 3rd degree polynomial fits, for active power and temperature, at 17:00 hours, using a 3-year weekday data

The higher degree polynomials deviate from the linear only for the extreme data-points and are, for the larger mid-section of temperature values, very similar to each other. The conclusion that follows is that the relationships do not deviate (extensively) from the linear (also discussed

in Section 4.5 with reference to the use of Pearson's and Spearman's correlation coefficients) and thus the relative "failure" of such models in predicting active power demand levels is due to multi-parametric effects (demonstrated by the correlation of the residuals of active power demand and temperature with other variables), which justifies the use of multiple regression analysis, particularly in situations where the desired outcome is accurate forecasting of active power demand levels.

Accordingly, multiple regression analysis (as discussed in Section 4.3) can be used to model active power demand with the use of two (or more) explanatory variables. An example with temperature and solar elevation angles is presented in Figure 4.38.

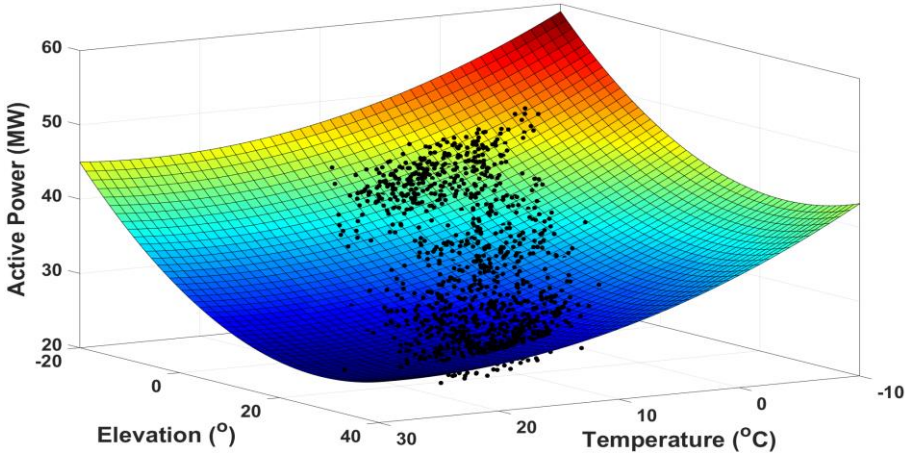


Figure 4.38: Example of multiple regression analysis: active power with temperature and solar elevation angles, at 17:00 hours, weekdays

The goodness-of-fit, compared with the simple linear regression of active power and temperature, for the particular half-hour of the day and for the dataset length described, increases from ~0.65 R^2 to ~0.9 R^2. It can also be demonstrated that the resulting residuals, of the multiple regression approach, have significantly reduced seasonal autocorrelations, shown for the actual values in Figure 4.39 (a) and for the sample autocorrelation in Figure 4.39 (b).

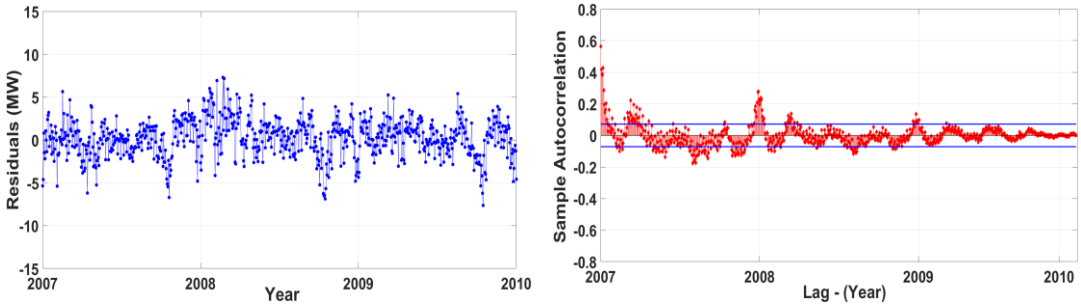


Figure 4.39: Residuals of multiple-regression analysis of active power with temperature and solar elevation angle, for 3-years at 17:00 hours: a) actual residuals (MW) and b) sample autocorrelation, (weekdays only)

Figure 4.40 shows the empirical and theoretical (Normal-Gaussian) distribution functions for three sets of residuals, i.e. residuals of the moving-average filter, residuals of the single-predictor regression model (active power with temperature) and the residuals of the multiple regression model (active power with temperature and solar elevation angle), all corresponding to the same active power dataset.

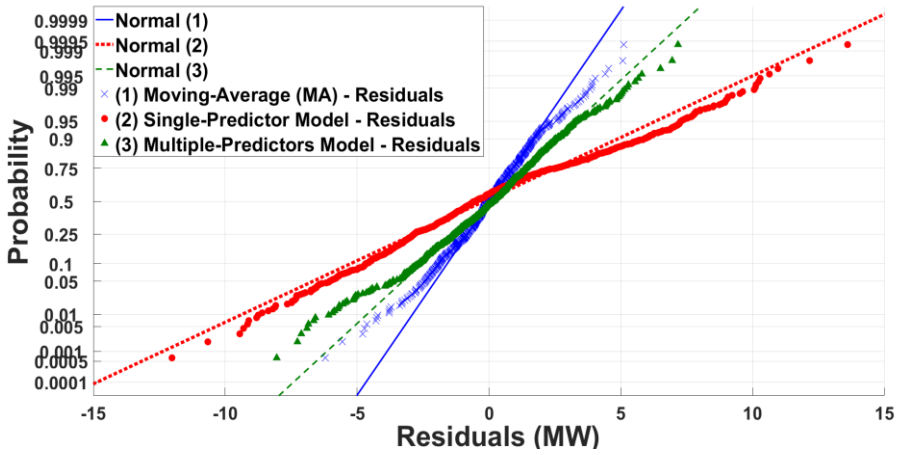


Figure 4.40: PP-plot for the residuals of the moving-average filter (± 10 weekdays), single-predictor and multiple-predictor models, with theoretical Gaussian distributions included

The range of the residuals is higher for the simple regression model and decreases for the multiple regression model, but still higher compared to the moving-average filter. Furthermore, the distribution of the residuals better fits the theoretical normal distribution for the multiple-regression model, excluding the values at the 5th and 95th probability range, whereas in the case of simple-regression, the distribution deviates from the theoretical for values above and below the 75th and 25th probability range, respectively. The results further support the assumption that multiple-predictor variable models have improved modelling performances (in the context of electricity demand modelling). The analysis presented in this section is provided for demonstrating some of the alternatives and extensions to the simple linear regression or simple, single-parameter, correlation analysis and the conclusions and methods described are used in the following chapters of this thesis. Specifically, the benefits associated with multiple-regression models are used for disaggregation, modelling and forecasting purposes in Chapters 6 and 7.

4.8 Moving-Window Regression and Seasonal Sensitivities

Moving-window regression, as described in Section 4.3, is adapted for the purposes of the methodology presented here, on a per half-hour of the day basis in order to determine the sensitivities of active power demand to the independent variables for smaller samples of the

datasets, that correspond to correlations within a moving-window throughout the duration of one calendar year. The approach can be summarised in the following steps:

- For all input variables, values are normalised with respect to the maximum of each dataset (3.3). Weekdays and weekends are separated, for reasons mentioned in the previous sections and in Chapter 3.
- For each pair of dependent/independent variables, for each half-hour of the day and for all available GSPs, the linear regression coefficients (R^2 and beta) are computed, for each weekday of the year within a moving-window of length ± 20 days. Each linear-fit is therefore determined using 41 data-points.
- For each coefficient and for each GSP, the resulting datasets have dimensions of 261 weekdays times 48 half-hours. For presentation purposes, the results shown are restricted to four characteristic hours of the day, i.e. 01:00, 08:00, 13:00 and 18:00 hours, for which the corresponding coefficients are calculated from the results from all available GSPs, in the form of the 50th percentile (median) values and for selected pairs of variables.

The selection of the pairs of dependent/independent variables is based on the results of the previous sections, primarily on whether strong correlations on the seasonal (per half-hour) time-frame have been established. The analysis is therefore concentrated on active power with: reactive power, temperature and solar elevation angles. Results for reactive power and voltage are excluded due to the, generally, weak correlations of these parameters with the meteorological and analemma variables. The selection of the four characteristic hours is based on an inspection of the final results, as well as on the results previously presented in Chapters 3 and 4, i.e. considering hours of the day with rapid rate of change in demands and periods of the day for which strong correlations have been established, in Sections 4.5 and 4.6.

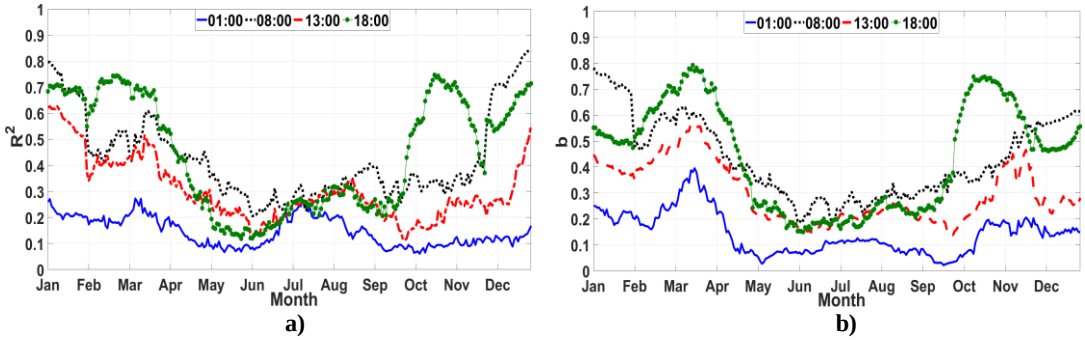


Figure 4.41: Moving-window linear regression results for active power with reactive power at characteristic hours of the day: a) R^2 and b) beta coefficients

Figure 4.41 shows the results of the moving-window linear regression analysis between active power and reactive power. Both coefficients (R^2 and beta) tend to be higher during winter

months and lower during summer months indicating higher/lower correlations between active and reactive power during the corresponding periods (on the day-to-day changes at particular half-hours). Peaks in R^2 and beta values are found, approximately, from February to the end of April and from October to the end of November, but values are relatively high between December and January as well. As discussed in Section 4.7, these two seasonal periods (February-April and October-November) are the ones for which the most extensive differences in the residuals of active power demand and temperature are found, i.e. most pronounced differences in demand levels for the same temperature levels. The current analysis shows that these are also the periods of higher common variability between active and reactive power demands, with respect to the day-to-day changes. Regarding the presented characteristic hours, the peaks are higher for 08:00 hours and 18:00 hours, with overall maximum coefficients shown for the latter.

An example of the results for a single GSP (GSP-14) is presented in Figure 4.42, using the R^2 and beta coefficients, in (a), as well as the normalised demands in (b), both corresponding to 18:00 hours. Periods with significant R^2 values are slightly shifted compared to the results for all GSPs in Figure 4.41 and are found between March to May and between mid-September to mid-November. GSP-14 is a predominantly residential substation and the yearly periods of rapid change in demand levels are more clearly evident, than when considering the full dataset, which contains GSPs of various customer-sector mixtures.

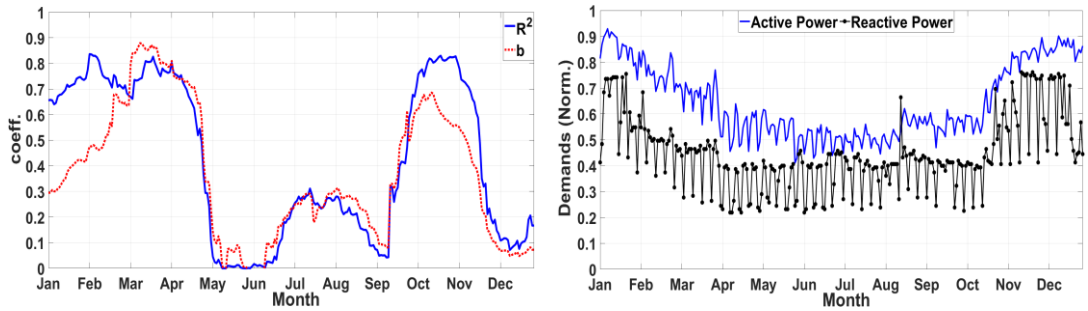


Figure 4.42: An example of the moving-window linear regression results for a single GSP at 18:00 hours: a) R^2 and beta coefficients and b) normalised demands for the same period

Figure 4.43 presents the results of the moving-window regression analysis for active power with temperature, for the characteristic hours mentioned above. Higher coefficient values are again found for the periods between February to April and October to November. With reference to the results presented in Section 4.7, the analysis presented here further supports the hypothesis that while the temperature levels are the same during the two periods, the discrepancies in demand levels are, at least to some extent, the result of psychological effects due to the corresponding positive and negative change in temperature levels, as well as

possibly due to the differences in solar irradiance levels that can affect lighting loads and occupancy levels. The beta coefficients are negative for the majority of the yearly period, but there is a clear tendency for the relationship to turn positive during the summer months, indicating increasing active power demands for increasing temperatures, which could potentially indicate the presence of AC loads. However, these beta coefficients correspond to best-fits with weak correlations, as shown in (a), for the R^2 values at the same yearly period.

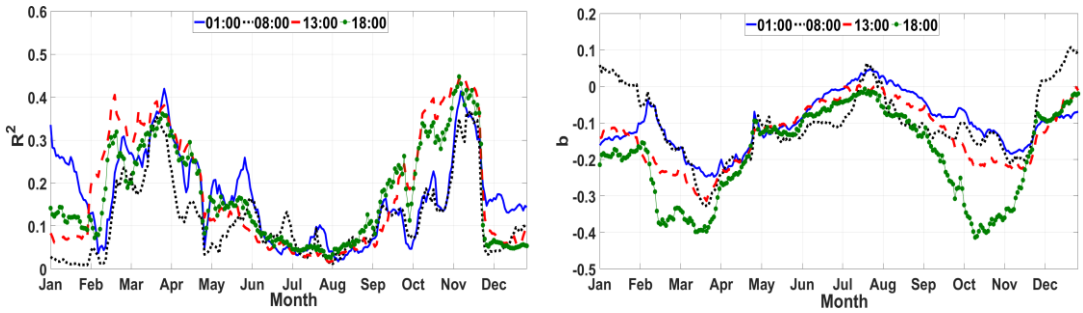


Figure 4.43: Moving-window linear regression results for active power with temperature at characteristic hours of the day: a) R^2 and b) beta coefficients

The change of the dependencies from positive to negative are better demonstrated by the comparison of the R^2 and beta coefficients between two different GSPs, one which is considered predominantly-residential, GSP-36 in Figure 4.44 (a) and a commercial/mixture, GSP-54, in Figure 4.44 (b).

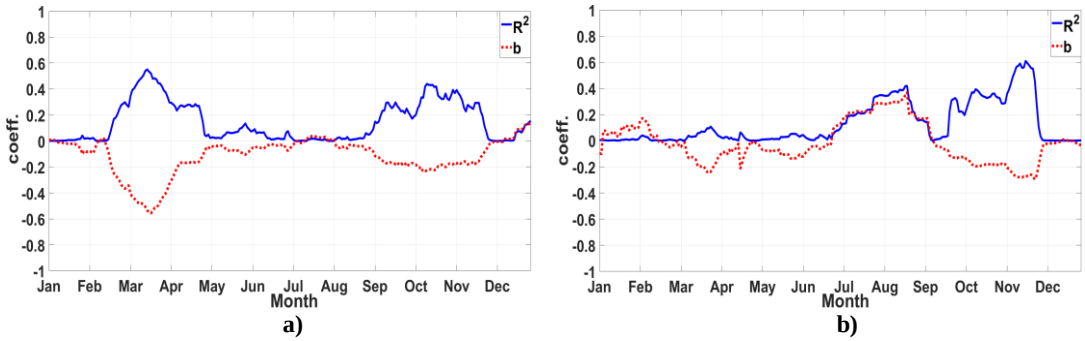


Figure 4.44: Comparison of the resulting coefficients for the moving-window regression analysis at 13:00 hours for: a) GSP-36 (residential) and b) GSP-54 (commercial/mixture)

The beta coefficients for the residential GSP are negative (or zero) throughout the year, while for the commercial GSP the coefficients turn positive for the summer period between mid-June and mid-September (during the same periods, there is also an increase in the corresponding R^2 values).

Figure 4.45 shows the results for the moving-window linear regression analysis between active power demand and solar elevation angles, for the four characteristic hours of the day. Similar patterns are noticeable but the increasing R^2 values between February-March and September-

November are more pronounced for the solar elevation angles. This is because, as previously mentioned, the analemma variables do not include day-to-day fluctuations and their seasonal components better correlate with active power demand changes.

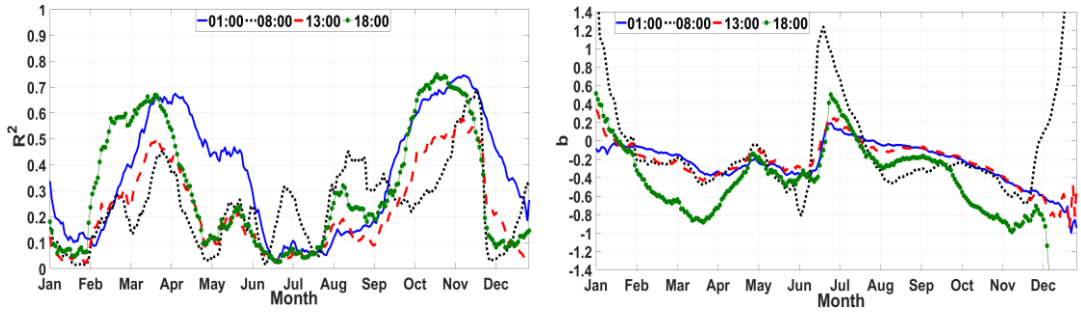


Figure 4.45: Moving-window linear regression results for active power with solar elevation angles at characteristic hours of the day: a) R^2 and b) beta coefficients

The beta coefficients are negative for the duration of the year and only turn positive for a short period during mid-June to mid-July (and to some extent during December and January) however these changes are not supported by high R^2 values, which indicates a very low goodness-of-fit for the linear-regression models calculated for the corresponding periods.

4.9 Chapter Conclusions

In the diurnal time-frame, correlations between active power and reactive power have been shown to be particularly strong; and consistently strong among the various GSPs for which *P* and *Q* data were available. In contrast, the seasonal and seasonal per half-hour *P*-*Q* correlations are relatively weak, compared to the diurnal correlations (though still moderate to strong), for the same datasets. This indicates that load composition changes more homogeneously, with respect to *P* and *Q* requirements, within single days, than through the year and it can be demonstrated using the standard deviations of the power factor (PF), as shown in Figure 4.46; based on the Scottish GSPs, for the seasonal PF variations at particular half-hours of the day, in (a) and for the daily PF variations within each day of the year, in (b).

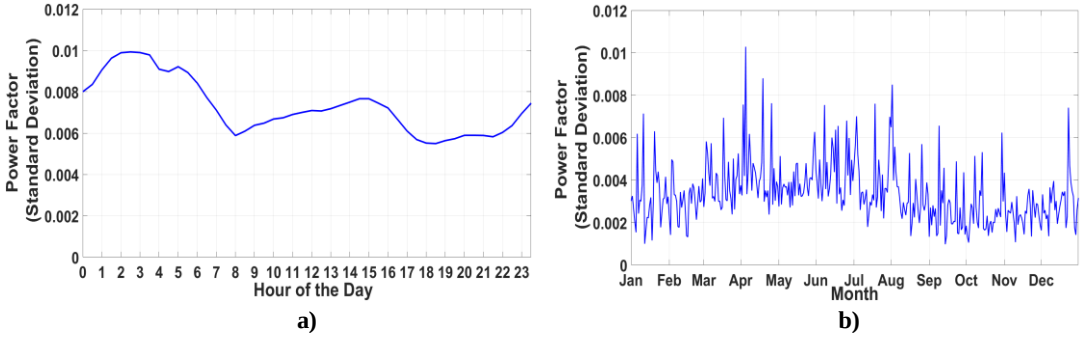


Figure 4.46: Standard deviations of power-factor (PF): a) through the year at particular half-hours and b) within each day, for all days of the year

As shown in Figure 4.46, the yearly standard deviations, in (a), are more than twice the range of the daily standard deviations, in (b). Furthermore, these are particularly high during night hours, i.e. increase in primarily-resistive heating loads, a fact which is further discussed in Chapter 6. In contrast with the *P*-*Q* relationship, the diurnal correlations of demands with the external weather and analemma parameters have been shown to be consequential of the similar daily patterns and cannot be considered as causally linked to increasing load-types, associated with changes in the corresponding external parameters (more precisely, such changes are not detectable in the diurnal time-frame).

In the seasonal perspective, active power is correlated with temperature and solar irradiance levels but more strongly with the analemma variables (particularly solar elevation angles), due to the fact that these are non-fluctuating, smooth representations of seasonality. Reactive power is correlated to a lesser extent with these parameters, indicating weaker dependence of *Q* to weather conditions which means that loads related to increase/decrease in reactive power are, mostly but not exclusively, determined by customer's schedules and seasonally fixed demand compositions.

The yearly periods which correspond to the highest rate of change of active power according to changing weather conditions, have also been established. These can be thought of as the periods of highest sensitivity to external parameters, concentrated during autumn and spring months, which are also the periods of highest rate of change of temperature levels. These changes cannot be accounted for by simple linear regression models (i.e. splitting of demands into upper and lower bands for smoothed values as shown in Figure 4.32) and require extensions to multiple-regression analysis for improved modelling performances, indicated by the extent and distribution of the residuals between the two approaches. In the same context, it has been demonstrated that modelling limitations are not a result of non-linear relationships, even though these are not absent and are expected to be more important when analysing data from different geographical locations.

Finally, regarding voltage, the analysis has failed to determine ways to usefully study statistical associations between voltage and active/reactive power as well as for voltage with external parameters. Voltage variations are shown to be, from a statistical perspective, mostly stochastic (also discussed in Chapter 3) and attempts to model these seasonal components based on filtered data result in voltage levels that can be considered "over-smoothed" and therefore not representative of the actual voltage variations.

The methodologies discussed and the conclusions drawn in this chapter, form the basis for subsequent analysis and particularly for the load-disaggregation and forecasting methods

presented in Chapters 6 and 7. The next chapter (Chapter-5), is concentrated on load-identification and GSP-classification, according to percentage contributions from different customer-sectors.

Chapter 5: Customer-Class Disaggregation

The results presented in Chapters 3 and 4 have demonstrated a diversity in the characteristic demand profiles, as well as in the dependency relationships between demands and external parameters, with respect to the total number of GSPs. The methodology presented in this chapter is based on the assumption that these differences are a result of load composition, which is primarily determined by the corresponding customer-class mixtures, at the individual GSPs, as expressed through demand levels in the diurnal and seasonal time-scales[14]. In most cases, aggregate demands at the MV-level correspond to contributions from the residential, commercial and industrial sectors (with additional loads from public areas, services and street lighting), each of which with its own percentage to the total measured consumption. The aim is, therefore, to determine these percentages for all available GSPs, based on the analysis of active power demand measurements.

The approach is based on the identification of specific patterns in measured demands, through the use of characteristic diurnal profiles and metrics that capture annual demand variability, all of which are presented in more detail in Section 5.1. In Section 5.2, an agglomerative clustering algorithm is used for the grouping of GSPs according to mean diurnal demand profiles, which are then arranged in order from primarily domestic to primarily non-domestic consumption, based on the assumed characteristic load profiles of the corresponding sectors.

The methodology is expanded in Section 5.3, through a more detailed customer-class disaggregation approach, utilising all available metrics from Section 5.1 and based on consumption statistics available from intermediate-geography-zone (IGZ) data. These are used as the training dataset, in order to identify patterns in the per half-hour metrics, that can be used to estimate percentages for the complete set of available GSPs. The results are presented according to the contributions from total residential (TR) and industrial and commercial (I&C) demands, as well as from sub-classes of the total residential sector corresponding to ordinary residential (OR) and economy-7 residential (E7) consumptions.

Therefore, in the context of this analysis, the following simplifications have been made: a) total measured demand can be separated into domestic (residential) and non-domestic (non-

[14] In fact, in Chapters 3 and 4 it has been shown that load composition (according to percentages of total residential demand, as identified in this chapter) correlates with load profiles (e.g. differences in normalised demands between weekdays and weekends – Section 3.4), as well as with different strengths of associations with weather conditions (e.g. regression analysis between active power and temperature – Section 4.6).

residential) demands, b) due to the format of the IGZ consumption statistics, industrial and commercial demands (I&C) are presented in a common category (although examples of predominantly industrial GSPs are also presented) and c) sub-classes corresponding to consumption from public areas/services/buildings (e.g. schools, hospitals, street-lighting), are included in the TR and I&C categories and are not investigated individually.

5.1 Metrics Used for the Analysis

Seven basic statistical parameters are calculated and subsequently used for purposes of clustering and customer-type identification/disaggregation. These are computed on a per half-hour of the day basis, using active power normalised values (3.3) and based on weekdays only. Therefore, for the complete diurnal period (i.e. 48 values), each parameter represents a particular profile, per GSP. For each of the seven statistical parameters a second normalisation is applied using the z-score values (3.1), resulting in a total number of 14 metrics, which are summarised in Table 5.1.

Table 5.1: Metrics used for clustering and customer-class disaggregation

No.	Metric (Norm.)	No.	Metric (Norm. and z-score)
1	$Mean_t$	***2***	z-score ($Mean_t$)
3	Max_t	***4***	z-score (Max_t)
5	Min_t	***6***	z-score (Min_t)
7	$Range_t$	***8***	z-score ($Range_t$)
9	$Range\_N_t$	***10***	z-score ($Range\_N_t$)
11	R^2_t	***12***	z-score (R^2_t)
13	$W\_Range\_N_t$	***14***	z-score ($W\_Range\_N_t$)

$Mean_t$ (Metric-1) is the per half-hour average measured active power over a one-year period, for each GSP. The z-score normalisation in Metric-2 and in all other cases, i.e. Metrics-4, 6, 8, 10, 12 and 14, are applied so that the diurnal profiles are adjusted to a common base, to allow for comparisons between GSPs with different demand levels. This has the additional benefit of highlighting different features in the diurnal profiles, as discussed in Chapter 3, Section 3.5. Max_t and Min_t, (Metrics-3 and 5) are calculated as the per half-hour 95th percentile and 5th percentile values (of normalised demands) over a one-year period. The per half-hour range, $Range_t$ (Metric-7), is defined as the difference between the yearly maximum and minimum values (i.e. $Max_t - Min_t$), while $Range\_N_t$ (Metric-9) corresponds to $Range_t$ divided by Max_t. This gives more weight to values that correspond to half-hours of lower maximum demand, as it is, for example, the case for the seasonal range during night-hours, in the presence of storage heating systems (i.e. there is more seasonal variability in demands, despite the fact that overall demand is lower than when compared to other periods of the day). R_t^2 (Metric-11) is the per half-hour coefficient of determination between active power and

temperature (as presented in Chapter 4, Section 4.6). Alternatively, R^2 coefficients for active power with solar elevation angles can be used, as the temperature and analemma variables are both good indicators of seasonality, as demonstrated in Chapter 4. Finally, $W\_Range\_N_t$ (Metric-13) is equivalent to Metric-9, but further weighed by the coefficient of determination in Metric-11, so that more weight is given to the seasonal range of variations, when this corresponds to a periodicity that matches the temperature seasonality.

The 14 metrics are illustrated in Figure 5.1, for a single GSP (GSP-14). Shown in (a) is the actual measured active power demand, in MW, for all weekdays of the year. In (b), the per-unit (3.3) normalised profiles are shown, which correspond to Metrics-1, 3, 5, 7, 9, 11, 13; and in (c) the z-score normalised profiles are shown, which correspond to Metrics-2, 4, 6, 8, 10, 12, 14.

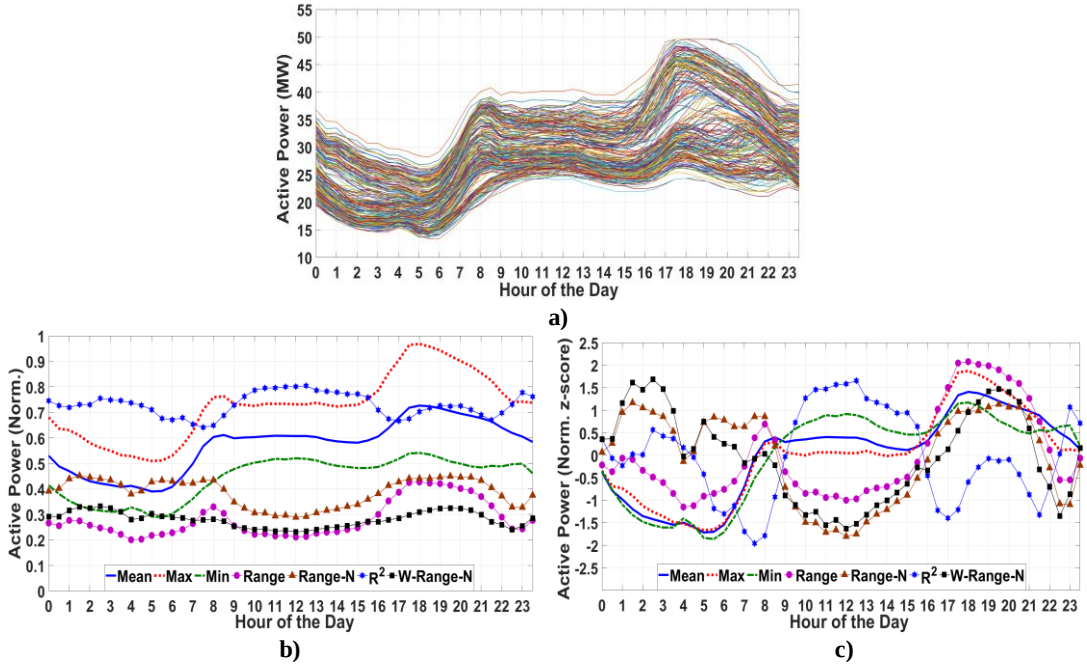
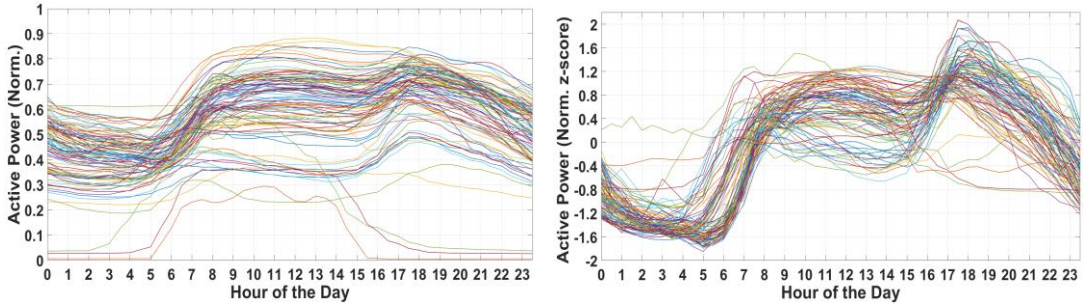


Figure 5.1: Example for GSP-14: a) original measurements (in MW), b) 7-metrics based on normalised values (3.3) and c) 7-metrics further normalised according to the z-score values (3.1)

The assumption, on which the subsequent analysis is based, is that since different customer-sectors (classes) are expected to have different consumption patterns, the identification of features based on a number of metrics and metric-normalisations, can be used to estimate the corresponding percentage-contributions. For example, in Figure 5.1 (b), the seasonal range, $Range_t$, peaks during the morning (i.e. from 07:00 to 09:00 hours) and during the afternoon to evening (i.e. from 16:00 to 22:00 hours), however for the normalised range, $Range\_N_t$, the highest levels extent during the night and particularly between 00:00 to 04:00 hours (this is also shown for $W\_Range\_N_t$). High levels of seasonal range during the night, are assume to

be related to economy-7 consumption. Similarly, the resulting diurnal patterns based on the z-score normalisation in Figure 5.1 (c), are more extenuated and at an equal level among the various GSPs and can, potentially, capture comparable differences, useful for the identification procedure. While the first normalisation, in Figure 5.1 (b), is sensitive to the maximum recorded demand at each GSP, the z-score values, in Figure 5.1 (c), will bring all profiles from all GSPs to the same mean level.

The differences between the two normalisation approaches are better illustrated in Figure 5.2, using all 98 GSPs; in (a) for Metric-1 ($Mean_t$) and in (b) for Metric-2 (z-score ($Mean_t$)).



a) **b)**

Figure 5.2: Diurnal profiles of 98 GSPs for a) $Mean_t$ and b) z-score ($Mean_t$)

Although the application of maximum value and z-score value normalisations modifies the original demand profiles in a way that they no longer correspond to actual consumption (in units of active power), they are better suited for the feature identification and extraction procedures, presented in the following sections.

5.2 GSP-Clustering

The GSP-clustering approach is based on an agglomerative algorithm, developed for the purposes of the presented analysis[15], by which the GSPs are arranged into groups according to their characteristic mean diurnal demand profiles (i.e. Metrics-1&2 from Section 5.1). This allows GSP-clustering irrespective of seasonal variability and levels of absolute demand, thus concentrating exclusively on the daily consumption patterns.

The resulting clusters (presented in Figure 5.5) are arranged into a descending order from predominantly residential to predominantly commercial and industrial demands (i.e. domestic and non-domestic loads) according to the final mean diurnal profiles of each cluster. The

[15] Although the clustering algorithm was developed for the purposes of this thesis, consultation of literature and of existing methods has shown that it is actually a form of agglomerative-hierarchical clustering, which are widely used for a number of different applications.

connection between profiles (or load shapes) and types of customers (i.e. sectors) is based on simple assumptions and available profiles from literature (e.g. Elexon profiles [89]). Specifically, diurnal profiles with mid-day peaks are assumed to be characteristic of non-domestic demands (particularly commercial, such as retail/offices and generally 09:00-17:00 working schedules), while profiles exhibiting morning and evening peaks are assumed to be related to domestic/residential demands. In Section 5.3 these clusters are evaluated against estimated percentages from contributions of: total residential (TR), industrial and commercial (I&C), ordinary residential (OR) and economy-7 residential (E7) demands.

The clustering procedure is initiated by calculating the pairwise similarities of mean diurnal demands, for all pairs of GSPs and quantified using the Pearson's correlation coefficient (4.2). The total number of GSP-pairs is given by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (5.1)$$

where n is the number of elements (i.e. 98 GPSs) and $k = 2$ for pairs, giving a total of 4753 correlation coefficients. Alternatively, different "distance" metrics can be used in order to quantify the pairwise similarities, including: Euclidean, cosine, Spearman's coefficient, city-block, etc. As the analysis is concerned with the similarities regarding diurnal patterns, irrespective of absolute demand levels, the Pearson's correlation coefficient is selected as an appropriate quantification (through extensive inspection of the results from various distance-metrics)[16]. The Pearson's coefficient, in contrast with the coefficient of determination R^2, may vary between [-1 1] which is useful because negative correlations can be regarded as showing increased dissimilarities.

The resulting pairwise coefficients are then arranged into a descending order and the GSPs belonging to the first pair (i.e. highest correlation coefficient) are assigned to a first common cluster. A similarity cut-off value, in terms of the Pearson's coefficient, i.e. r_{co}, is selected, in order to specify the level of similarity for which two GSPs are assigned to a common cluster. The algorithm is executed in iteration steps of increasing r_{co} values ([0 1] in steps of 0.01) and a final decision on the desired level of similarity is made, after inspection of the resulting clusters at each level.

[16] This is not to imply that other approaches were unsuccessful. In fact, Spearman's coefficients give, approximately, the same results and when using z-score normalised values, Euclidean distances also give very similar results.

In each iteration step and provided there is a non-zero number of GSPs not yet assigned to a cluster, the procedure continues assigning GSPs to clusters based on:

I. the pairwise correlation coefficient and

II. a second correlation coefficient as calculated between un-assigned GSPs and the mean profiles of existing clusters.

The procedure, at each cut-off level - r_{co}, is terminated when all GSPs have been assigned to a specific cluster. If, for any GSP, the criteria are not met, i.e. none of the resulting correlation coefficients (I or II, from above) is higher than the cut-off level, then the corresponding GSP is assigned to its own cluster. A schematic representation of the algorithm is shown in Figure 5.6. The number of clusters at each cut-off level are presented in Figure 5.3, as estimated using Metrics-1&2 in (a). For comparison, the number of clusters resulting from the use of different metrics (from Table 5.1) are shown in (b).

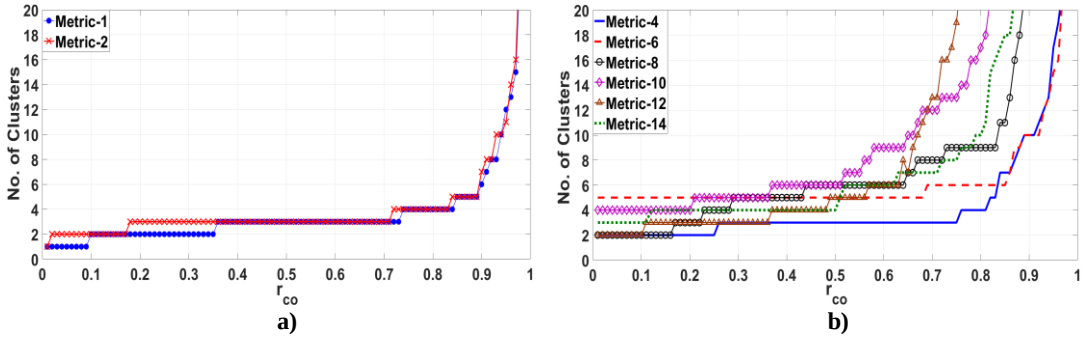


Figure 5.3: Number of resulting clusters per similarity cut-off values r_{co}, for mean diurnal demand profiles based on a) Metrics-1&2 and b) Metrics-2,6,8,10,12,14

Metrics-1&2, in Figure 5.3 (a), produce almost identical number of clusters for the various cut-off levels, with the z-score normalised profiles (Metric-2) showing slightly faster increase at the lower r_{co} levels. In Figure 5.3 (b) the number of clusters for the rest of the metrics are presented (for convenience only the z-score normalised metrics are shown, i.e. even numbers from Table 5.1). These tend to increase rapidly from very low cut-off levels and particularly for Metrics-8, 10 ,12 and 14. The number of clusters for sufficiently high cut-off levels is maintained low only for Metrics-4&6, which correspond to the Max_t and Min_t diurnal profiles.

Therefore, although classification is possible based on Metrics-4 to 14, for low cut-off values, GSPs in common clusters would include non-similar diurnal profiles (i.e. r_{co} is very low to make useful distinctions between the profiles), whereas higher cut-off levels result in a large number of clusters. This means that the commonality of features among GSPs is considerably low for the corresponding metrics and thus the results can be considered as a poor clustering-

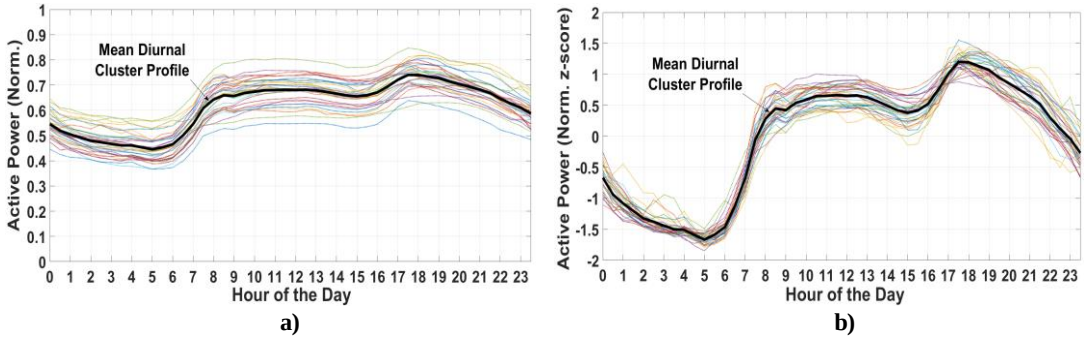
classification (e.g. at cut-off $r_{co} = 0.9$, Metric-8 gives 23 clusters which is ~25 % of the total number of GSPs).

The results presented in this section are based on a similarity cut-off level of $r_{co} = 0.95$ and for clustering performed on mean diurnal profiles, as expressed by Metric-2. This arrangement produces a total number of 11 clusters, which are shown in Table 5.2, with the corresponding GSP-numbers belonging to each cluster. These clusters are also arranged in an order of (assumed) change from predominantly residential (C-1) to predominantly commercial (C-6) GSPs, as well as characteristic demands that can be attributed to industrial (or other/atypical) loads, as shown for C-7 (these have diurnal profiles and peak demands that cannot be categorised according to the simple assumptions of mid-day and morning/afternoon peaks, discussed before). Furthermore, 4 out of 98 GSPs (~4 %) are assigned to unique clusters, as shown in Table 5.2, for Cluster No. 8, 9, 10 and 11, due to the fact that they could not be grouped based on the selected similarity cut-off level.

Table 5.2: GSPs per cluster, for a similarly cut-off level of $r_{co} = 0.95$

Cluster No.	No. of GSPs	GSPs No.
1	11	78, 79, 82, 83, 84, 85, 87, 90, 92, 95, 97
2	5	15, 17, 47, 48, 50
3	10	59, 61, 62, 64, 65, 66, 68, 76, 81, 98
4	33	4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 18, 20, 21, 22, 23, 25, 27, 28, 29, 32, 33, 35, 36, 37, 38, 39, 40, 43, 45, 46, 49, 71, 73
5	11	55, 56, 57, 67, 69, 72, 74, 75, 77, 86, 88
6	21	1, 2, 3, 9, 16, 24, 26, 30, 31, 34, 41, 42, 44, 51, 52, 53, 54, 58, 60, 63, 94
7	3	89, 91, 93
8,9,10,11	1 (each)	70, 96, 80, 19 (accord. to column 1)

Mean diurnal cluster profiles can be calculated as the average per half-hour values of all its member GSPs, as shown in an example for Cluster-4, in Figure 5.4.



a) **b)**

Figure 5.4: Cluster-4 mean diurnal profile and corresponding GSPs: a) for normalised values and b) for z-score normalised values

Figure 5.4 (a) is based on the normalised mean diurnal profiles (Metric-1), while (b) is based on the z-score normalised profiles (Metric-2). Note that, as shown in Figure 5.3, at this r_{co} level, the two metrics give the same number of clusters and these clusters are in fact populated by the same GSPs. The mean diurnal profiles of the first seven clusters are presented in Figure 5.5 (excluding the single-GSP clusters, i.e. C-8, 9, 10 and 11).

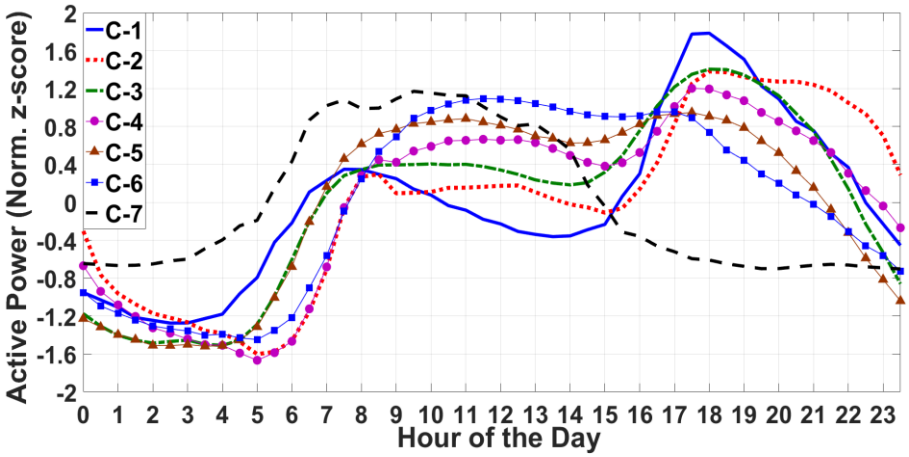


Figure 5.5: Mean diurnal profiles of the first 7 clusters, for a similarity cut-off level of 0.95

There are periods of the day when these profiles are very similar/close to each other, particularly between 00:00 to 03:00 hours. The most prominent differences in the diurnal features are the morning, mid-day and evening peaks, which are at different levels for the resulting clusters, showing extended morning and evening demands for the predominantly residential GSPs, in C-1 and C-2. The pattern starts to shift for C-3 and C-4 with a noticeable increase in mid-day consumption and for C-5 mid-day demands are at, approximately, the same level as the afternoon/evening peak, whereas in C-6 there is a clear mid-day peak between 11:00 to 12:00 hours.

While the presented classification approach and the resulting clusters give diurnal profiles that can be considered as reasonable estimations of various mixes of customer-classes, it should be noted that the final step of the analysis is not automated, nor does it provide detailed customer-class disaggregation. The ordering of the resulting clusters, in Table 5.2 and Figure 5.5, with respect to domestic (residential) and non-domestic (commercial & industrial) demands is performed manually and based on assumed characteristic profiles, which are not necessarily correct for the individual GSPs in each cluster. A more analytical procedure for customer-class identification and disaggregation is provided in the next section, where the classification performance of the current section is also illustrated.

Note: the next page presents the clustering algorithm used in the current section.

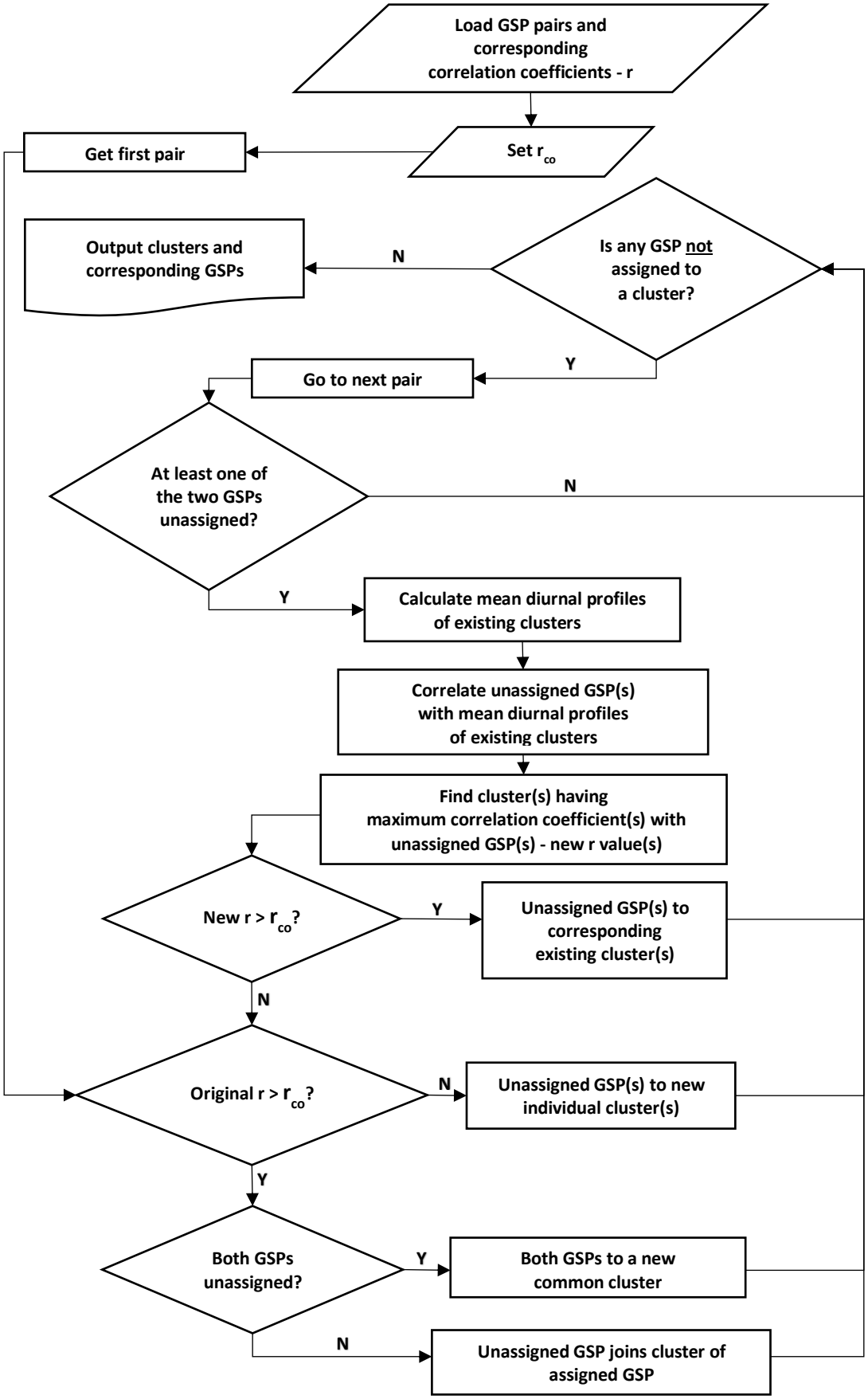


Figure 5.6: Clustering algorithm

5.3 Customer-Class Disaggregation

5.3.1 Overview of Approach

The 14-metrics presented in Section 5.1 are used to construct combinations-of-metrics (CoM), i.e. products and ratios between metrics corresponding to different periods of the day, which are considered as the independent variables. The dependent variables correspond to a limited number of percentage-contributions, for four customer-classes, i.e. total residential (TR), industrial and commercial (I&C), ordinary residential (OR) and economy-7 residential (E7), which are estimated from available sub-national consumption information, for 11 GSPs and are referred to as the target dataset.

The two datasets, i.e. independent and dependent (target) variables, are used as inputs in an exploratory regression analysis, based on simple linear-regression as described in Chapter 4, Section 4.3.2 and the goodness-of-fit is quantified using the coefficient of determination, R^2. The purpose is to identify CoMs, with sufficiently high R^2 values, that can be used to determine the corresponding percentage-contributions for all four customer-classes and for the full number of available GSPs (i.e. 98).

Initially, the analysis is performed to the highest available data resolution (i.e. 48 half-hours) and therefore the total number of tested models (i.e. CoMs), is given by:

$$(No_t)^2 \times (No_{Metrics})^2 \times No_{Operations} = No_{Models} \quad (5.2)$$

where No_t is equal to 48 half-hours, $No_{Metrics}$ is equal to 14 (from Section 5.1), $No_{Operations}$ is equal to 2, for multiplication and division and No_{Models} is therefore equal to 903168. In each case and depending on the selected operation, the independent variable X_i (i.e. CoM), is given by:

$$X_i = M_A(t_c) \cdot M_B(t_d) \quad (5.3)$$

where X_i is the i^{th} product (or ratio) between metrics M_A at half-hour t_c and M_B at half-hour t_d. Note that the approach allows for combinations such that $\mathrm{M_A} = M_B$, as well as for $t_c = t_d$. The methodology is therefore based on the assumption that products (or ratios) of metrics between different diurnal periods can be linearly correlated with demand percentages from various customer-classes and for the selected GSPs for which target data is available. The assumption is supported by the results in Section 5.2 which showed that the distinctions between clusters (Figure 5.5) are particularly pronounced at certain diurnal periods, i.e. morning, mid-day and evening peaks. However, considering the high number of tested models, i.e. 903168, and the small size of the target-dataset (described in detail in the next section), it

is expected that some high R^2 values (indicating satisfactory model predictive power) will be acquired purely by chance. This is reinforced by the use of various metrics and metric-normalisations, which may result in more spurious correlations.

Therefore, the presented methodology does not rely on a single model (CoM) for percentage estimations, nor for validation, but rather on the overall occurrence of best combinations of metrics at specific half-hours of the day. Furthermore, and in order to address these issues, it is subsequently generalised into a diurnal-blocks approach, which does not consider each half-hour of the day as an individual degree of freedom, but rather uses the resulting patterns of best CoM occurrence to construct four diurnal-blocks (these are described in the next section and are shown in Figure 5.12). The diurnal-block approach reduces the total number of models to 6272, according to:

$$(No_{BL})^2 \times (No_{Metrics})^2 \times No_{Operations} = No_{Models} \quad (5.4)$$

which is equivalent to (5.2) but $(\mathrm{No_t})^2$ is now substituted by $(\mathrm{No_{BL}})^2$, where $No_{BL} = 4$. In this case, the independent variables, i.e. CoMs, are given by:

$$X_i = M_A(BL_c) \cdot M_B(BL_d) \quad (5.5)$$

where X_i is the i^{th} product (or ratio) between metrics M_A at block BL_c and M_B at block BL_d. M_A and M_B are given as the average values of the corresponding metrics within the window defined by the half-hours in each block. The exact range of each of the four blocks within the diurnal-period are presented in Section 5.3.3, as these are determined based on the results of the per half-hour analysis. The final decision for which CoMs should be used for the estimation of percentages for the four customer-classes and for all GSPs is a matter of determining the consistency of the models of best performance, over both the per half-hour and the diurnal-blocks approaches. This discussion and the corresponding analysis are presented in Sections 5.3.3 and 5.3.4. The next section, Section 5.3.2, describes the process for determining the percentages for the target-dataset.

5.3.2 Target Datasets

Regarding the target percentages, i.e. dependent variables; due to the absence of available information on TR, I&C, OR and E7 for individual GSPs, the following procedure is used to calculate them, based on the domestic and non-domestic electricity estimates (2009) regarding intermediate-geography-zones (IGZs) in Scotland [128], [212]:

- A number of Scottish GSPs are identified using maps provided from the corresponding DNO [213] as well as using google maps. The DNO maps show the interconnections of

primary/secondary substations and allow for a more accurate identification of the grid topology, while the satellite images allow for a clearer identification of the inhabited areas, which are of close proximity to the corresponding substations. An example of an identified GSP and the relevant areas is shown in Figure 5.7, for GSP-14 (note that this is a simplified example of the area identification procedure, as only the satellite image is shown).

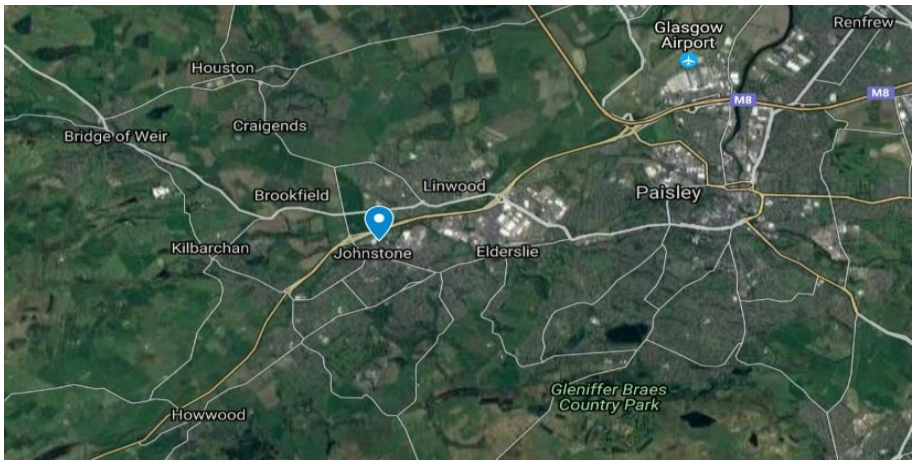


Figure 5.7: GSP-14 and corresponding supplied areas

- The selected areas are associated with their respective IGZs, for which there is available data regarding total consumption from TR, I&C, OR and E7 customers, based on the aggregated measurements from customer electricity meters or meter point administration numbers (MPANs). An example of the linkage between IGZ-codes and the corresponding areas is presented in Figure 5.8, for total residential (domestic) consumption, which is comprised of ordinary (OR) and economy-7 (E7) consumptions. For Scotland, the intermediate-geography-zones are contained within council areas, or local authorities, indicated by the LA-name in Figure 5.8 (b) and the IGZ-codes, in Figure 5.8 (a), are selected from the satellite and DNO maps, as described above.

S02000983	Lochwinnoch
S02000984	Renfrewshire Rural South, Hawkhead and Howwood
S02000985	Paisley Glenburn West
S02000986	Paisley Glenburn East
S02000987	Paisley Foxbar
S02000988	Johnstone South West
S02000989	Paisley Dykebar
S02000990	Paisley South West
S02000991	Paisley South East
S02000992	Johnstone South East
S02000993	Paisley South
S02000994	Johnstone North West

a)

LA name	LA code	IGZ code	Ordinary domestic consumption	Economy 7 consumption
Renfrewshire	UKM3502	S02000987	7,245,450	2480107.4
Renfrewshire	UKM3502	S02000988	7,411,122	2622450.3
Renfrewshire	UKM3502	S02000989	5,181,445	745210.5
Renfrewshire	UKM3502	S02000990	7,912,094	1882656.4
Renfrewshire	UKM3502	S02000991	8,340,557	3292363.4
Renfrewshire	UKM3502	S02000992	6,144,138	1984781.2
Renfrewshire	UKM3502	S02000993	6,504,005	1093643.7

b)

Figure 5.8: a) IGZ-codes and b) corresponding domestic consumption (kWh)

- The total consumption, per selected GSP, is calculated by adding all yearly consumptions from the corresponding/identified areas from all customer-classes. Then, total

consumption for each of the four customer-classes is used to calculate their individual percentage-contributions.

Note, that due to the format of the available sub-national statistics, total consumption consists of total residential (TR) and industrial and commercial (I&C) percentages, such that:

$$Total\ Consumption\ (\%) = TR\ (\%) + I\&C\ (\%) \qquad (5.6)$$

and, similarly, total residential (TR) consumption consists of percentages of ordinary residential (OR) and ecnomy-7 (E7) residential, i.e.:

$$TR\ (\%) = OR\ (\%) + E7\ (\%) \qquad (5.7)$$

It should be noted that (5.6) and (5.7) are necessary simplifications. In reality, the methodology, as described by the authors (Department for Business, Energy & Industrial Strategy in [212]), has limitations and there are portion of unallocated consumption. However, these are given for the LA-areas and not for the IGZ-areas and therefore they could not be included in the current analysis.

The accuracy of the target-datasets can be investigated by comparing the total consumption, as estimated from the IGZ data, with the total consumption as calculated from the active power demand data, per GSP. The results are presented in Figure 5.9. Figure 5.9 (a) shows the linear fit between the two consumption estimations and Figure 5.9 (b) shows the percentage error (with respect to the measured, active power demands). There is, generally, a very good agreement between the two, with the error being kept below 16 % for all GSPs and below 5 % for 5 out of 11 of the GSPs.

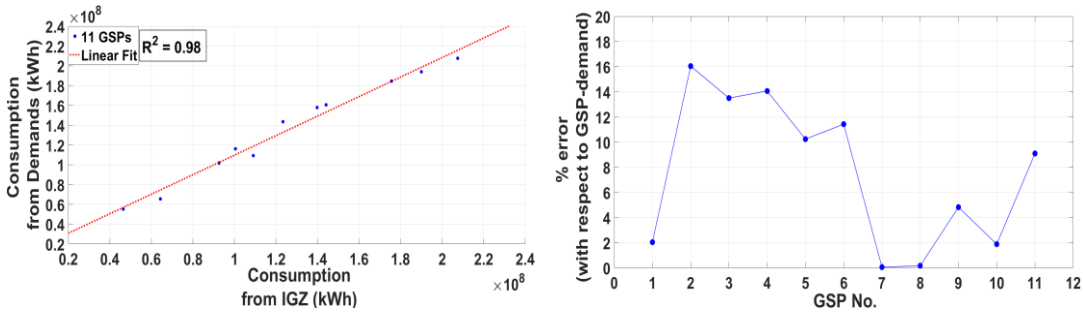


Figure 5.9: a) Estimated consumption from IGZ-data vs consumption calculated from active power demands and b) percentage error

The procedure for the target data estimations cannot be used as a generalised approach for customer-class disaggregation for a number or reasons. Firstly, each IGZ is not explicitly associated with a specific GSP and determining which areas are supplied by which GSP was, for the results presented here, a matter of manual inspection of the area maps. Furthermore,

each GSP does not necessarily supply a single IGZ and vice-versa. There is, therefore, a large number of intersections and the interconnected nature of the distribution grid means that there are no clear boundaries for determining the exact demands from single GSPs at individual IGZs. This is the reason why the target dataset is limited to 11 out of 98 GSPs (or ~10 % of the sample). The target dataset is also limited to Scottish GSPs only, which means that the resulting estimations may not be representative of the demand composition from different geographical locations with different grid and consumption characteristics.

5.3.3 Model Training and Optimal Models

Figure 5.10 shows the distribution of successful CoMs (5.3) over the 48 half-hours of the day for the operations of division and multiplication. Figure 5.10 (a) shows the results for (TR) and (I&C), which are equivalent for the occurrence of successful CoMs, as described by (5.6). The periods of occurrence of successful CoMs for the OR and E7 classes are presented in Figures 5.10 (b) and (c), respectively. Occurrence (y-axis), is relative to the number of linear-fits with an R^2 value above some arbitrarily chosen limit. For the analysis presented here, the limits are set to: $R^2 \geq 0.9$ for TR/I&C and OR and $R^2 \geq 0.8$ for the E7 (due to the poorer overall performance of the E7 models).

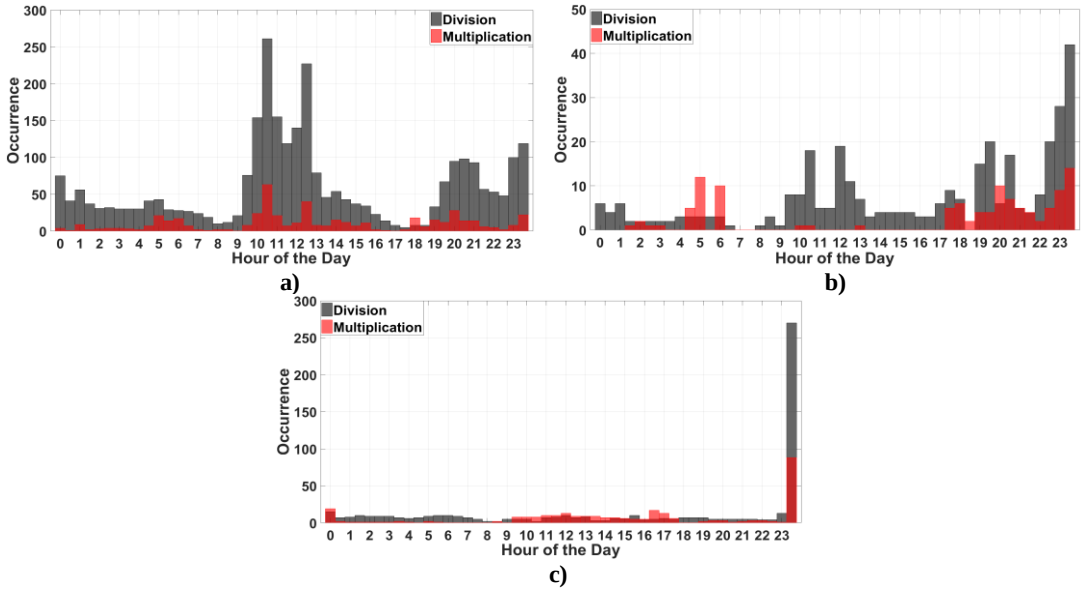


Figure 5.10: Distribution of successful metrics among the half-hours of the day for: a) TR and I&C, b) OR and c) E7

There is, generally, a lower number of successful combinations-of-metrics for the operation of multiplication, compared to division, indicating that ratios of metrics at different half-hours of the day are more suitable. For TR consumption, values of t_c and t_d (5.3) are concentrated during the mid-day period, i.e. between 10:00 to 13:00 hours, as well as between 19:00 to

00:30 hours. Similar results are shown for OR consumption, but for a lower number of successful CoMs and with a distinctive peak at 23:30 hours. This is also the only period with a high number of successful CoMs for E7, as shown in Figure 5.10 (c). The results are in agreement with the basic assumption that mid-day and evening peaks are good predictors of TR (and therefore I&C) consumption. The 23:30 peaks for OR and E7 are also justifiable based on the fact that this is, approximately, the period that signifies the start of night-hour tariffs for the economy-7 meters and it is therefore the period when demand features are more suitable for identifying the differences between the two customer-classes.

Figure 5.11 shows the corresponding successful metrics M_A and M_B, for TR (and I&C) in (a), OR in (b) and E7 in (c). This, as in Figure 5.10, corresponds to the frequency of occurrence of models with a coefficient of determination higher than the selected thresholds (i.e. $R^2 \geq 0.9$ and $R^2 \geq 0.8$). While Figure 5.10 showed the periods for which the metrics correspond to, Figure 5.11 shows the frequency of occurrence of the metrics themselves, in the set of best CoMs (as defined by the selected R^2 thresholds).

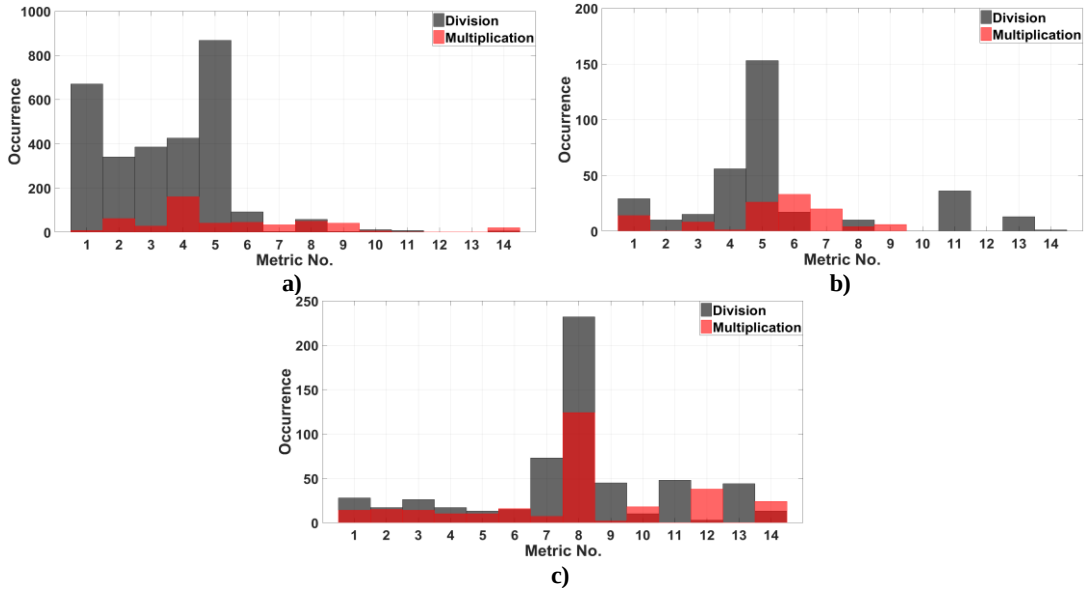


Figure 5.11: Successful metrics for: a) TR and I&C, b) OR and c) E7

For TR, predictability is higher for the first five metrics and particularly for Metrics-1 and 5, i.e. mean and minimum diurnal profiles. Metric-5 is also the peak for OR, followed by Metric-4 (z-score normalised maximum) and Metric-11 (per half-hour R^2 between active power and temperature, from Chapter 4). Predictability for the E7 peaks for Metrics-7 and 8, i.e. normalised and z-score normalised range of variations. This is also justifiable based on the assumption that seasonal variations for GSPs with a significant number of E7 meters can be

associated with thermal heating demand, particularly during night hours, or, as shown in Figure 5.10, at the period of commencement of economy-7 night tariffs.

Table 5.3 presents the optimal (highest R^2 values) CoMs, at specific half-hours of the day. The three combinations with the best performances are shown, for each customer-class and for each operation (i.e. multiplication and division). The total number of "successful" combinations, with respect to the selected limits (i.e. $R^2 \geq 0.9$ and $R^2 \geq 0.8$) are: 274 and 1435 for TR/I&C, 56 and 170 for OR and 146 and 292 for E7. In the case of multiplication, the duplicate values have been discarded (these result from the fact that multiplication is a commutative operator, i.e. $a \times b = b \times a$).

Table 5.3: Optimal CoMs for the per half-hour analysis

Customer Sector	Operation	M_A	M_B	t_c	t_d	R^2
TR and I&C	×	3	6	23:30	06:00	0.97
		8	14	10:30	11:00	0.96
		4	6	05:30	05:00	0.96
	÷	1	1	10:30	23:30	0.97
		2	5	10:30	04:00	0.97
		2	5	10:00	04:00	0.97
OR	×	6	6	01:30	05:00	0.95
		5	7	23:30	20:00	0.95
		5	7	23:00	18:00	0.94
	÷	5	5	12:00	23:30	0.97
		5	5	11:30	23:30	0.96
		5	5	12:30	23:30	0.96
E7	×	8	12	12:30	17:00	0.92
		8	12	12:30	16:30	0.91
		8	12	13:00	16:30	0.91
	÷	9	7	23:30	15:00	0.94
		7	9	15:00	23:30	0.93
		7	9	12:30	23:30	0.92

Table 5.3 shows better consistency for the OR and E7 combinations with respect to the number of metrics used and the corresponding half-hours. TR includes a larger variety of metrics (Metric-3, 6, 8, 14, 4, 1, 2 and 5) indicating that the estimation of total residential consumtpion is possible through the distinctions of more diurnal features, than for the rest of the customer-classes. This is also reflected in Figures 5.10 and 5.11, which showed more diversity in the distribution of metrics among half-hours, as well as among the succesful metrics, for TR demands.

The limitations previously discussed, regarding the small number of GSPs in the target dataset are represented by the very high coefficients for the optimal combinations (above 0.9 in all cases). It is expected that a similar analysis with a larger target dataset would result in an overall decrease in the goodness-of-fit for the corresponding models, due to the increased

probability for the presence of GSPs that to not adhere to the similar demand patterns[17]. In the same context, since the analysis is flexible up to the highest diurnal resolution (48 half-hours), the optimal results of Table 5.3 may be considered too specific for the input data and therefore more prone to erroneous estimations for GSPs of different consumption characterstics[17]. It is, however, possible that for the OR and particularly for the E7 categories, metric combinations at individual half-hours the day are more appropriate (than the diurnal-block analysis, which is presented in the following pages) because these can capture finer, detailed demand patterns and produce more accurate estimations. The large number of models in (5.2) also increases the computation time, although this does not exceed ~30 minutes, based on processing implemented on the PC-desktop system described in Chapter 1.

As mentioned in Section 5.3.1, and in an attempt to address some of the issues discussed, the methodology is modified so that No_t (5.2 and 5.3) is replaced by No_{BL} (5.4 and 5.5), where *BL* corresponds to four diurnal-blocks and the total number of models is reduced to 6272. The diurnal-blocks are determined after examination of the results in Figure 5.10, i.e. by considering sets of half-hours of the day (periods) for which model performance is increased, as well as from inspection of the results in Table 5.3, so that t_c and t_d are not included as elements of the same block. To improve the block selection process, the variance of the metrics among all 98 GSPs is also considered, on a per half-hour of the day basis and quantified as the standard deviation, shown in Figure 5.12.

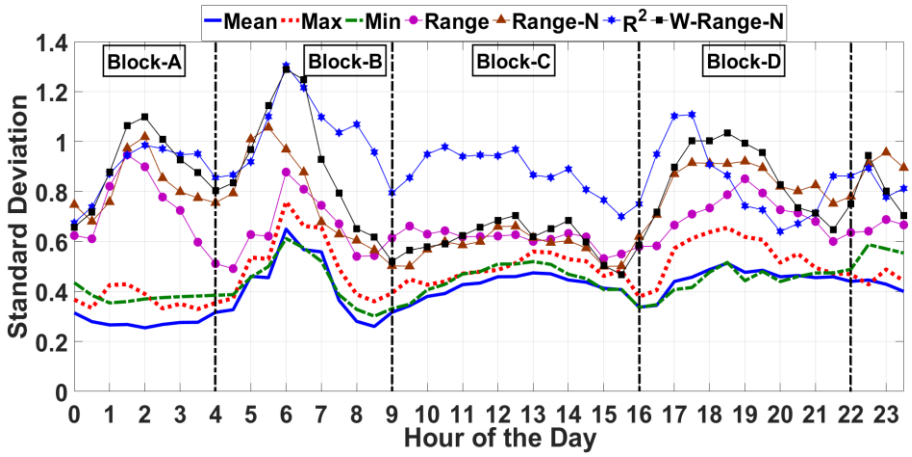


Figure 5.12: Variability of metrics (in standard deviation) among 98 GSPs and the 4-selected diurnal-blocks

[17] This is not intended to imply that the results are considered inaccurate. In fact, the resulting combinations of metrics and their periods of occurrence make sense from a theoretical perspective, i.e. the expected patterns are the ones providing best model performances (e.g. range of variations for E7; mid-day demands for TR; etc.). However, a larger data-sample would be able to train the models for a larger variate of consumption characteristics and would also add more statistical validity.

The results in Figure 5.12 are based on the analysis of the z-score normalised metrics, so that the per half-hour variability among GSPs is restricted to differences in diurnal patterns and not to differences in actual demand levels (as discussed for Figure 5.2). The selected diurnal blocks are: a) Block-A between 22:00 to 04:00 hours, b) Block-B between 04:00 to 09:00 hours, c) Block-C between 09:00 to 16:00 hours and d) Block-D between 16:00 to 22:00 hours. The combinations-of-metrics, i.e. independent variables, previously given by (5.3), are now modified in terms of blocks and calculated according to (5.5).

Figure 5.13 shows the distribution of the successful CoMs over the four diurnal blocks for a) TR (and I&C), b) OR and c) E7 consumption. In the diurnal-blocks analysis, the limits of what is considered a satisfactory linear fit are reduced and the new thresholds are defined as: $R^2 \geq 0.8$ for TR/I&C and OR and $R^2 \geq 0.7$ for the E7.

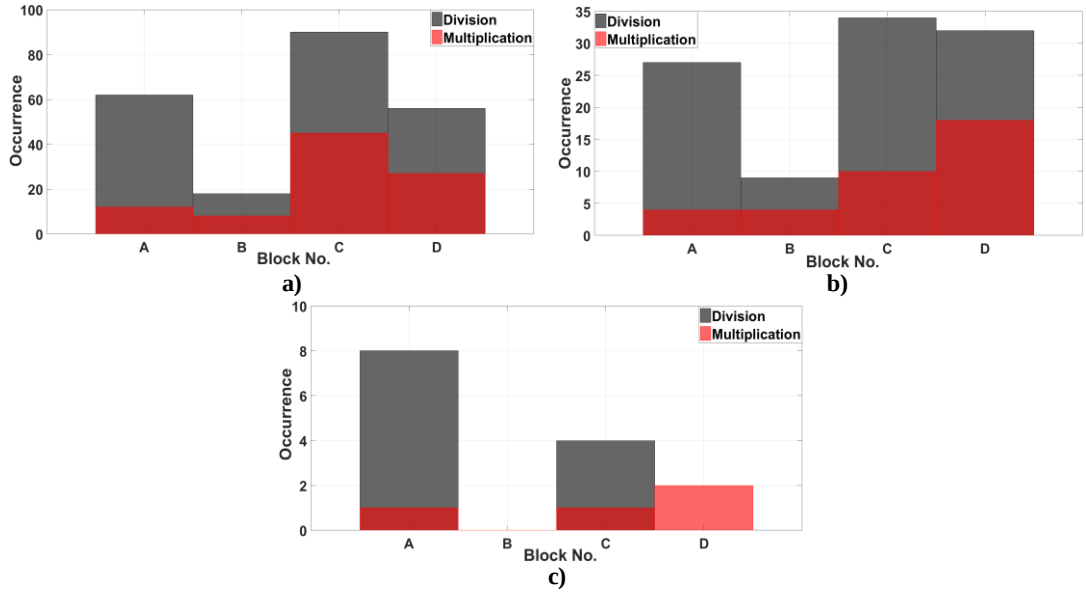


Figure 5.13: Distribution of successful metrics among the four diurnal-blocks for: a) TR and I&C, b) OR and c) E7

Interestingly, Block-B which shows the highest variability of metrics among the 98-GSPs (Figure 5.12), has only a limited number of occurrences in the successful CoMs. This is also reflected in the occurrence of successful metrics as presented in Figure 5.10 and it implies that demand variability (among GSPs), during the early morning hours cannot be associated with percentages from different customer-classes. A possible explanation is that during that period, both domestic and non-domestic sectors experience and increase in electricity demands and therefore the features are not suitable for distinguishing between the two (unless if done with respect to weekday/weekend distinctions, as demonstrated in Chapter 3, Figure 3.28). For TR and OR consumption, the estimations most frequently include Blocks-A, C and D while for the E7 consumption, combinations of metrics are shown for Blocks-A and C. The total number

of successful combinations is very low for the E7 consumption (i.e. 16) even at the reduced R^2 threshold value of 0.7.

Figure 5.14 shows the corresponding successful metrics M_A and M_B, for TR (and I&C) in (a), OR in (b) and E7 in (c). As in the case of the per half-hour analysis, while the previous figure presented the periods of occurrence of successful CoMs, Figure 5.14, shows the metrics that constitute the successful CoMs.

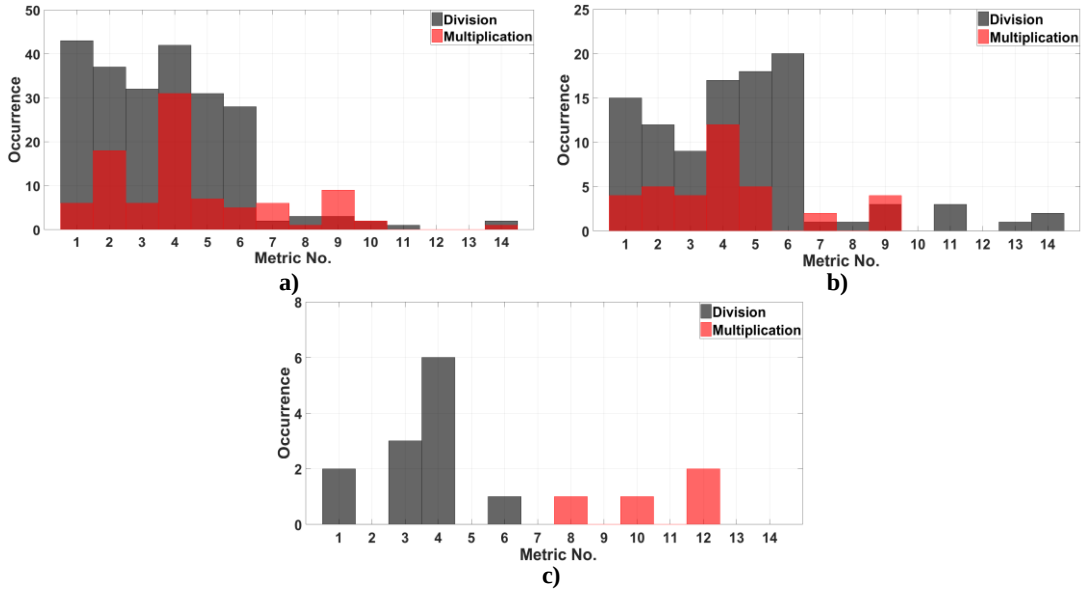


Figure 5.14: Successful metrics for: a) TR and I&C, b) OR and c) E7, for the diurnal-blocks analysis

For TR, the results are similar to the per half-hour analysis (Figure 5.11), there is however an increase in the occurrence of CoMs including Metric-6, while the peaks shown in Figure 5.11 for Metrics-1 and 5 are not as distinctive for the diurnal-blocks analysis (in fact, peaks are now shown for Metrics-1 and 4). There is also an increase in the variety of successful metrics for the OR estimations, which include Metrics-1 to 6, as in the case of TR. These results indicate that the block generalisation produces more combinations that have potential applicability, at least for the TR and OR categories. For E7, peaks are shown for Metrics-3 and 4 and only a marginal success for Metrics-1, 8, 9 and 12. The overall success of the E7 models is reduced, thus the adjustment of the threshold values to $R^2 \geq 0.7$. This implies that, as mentioned before, for the E7 estimations, the initial higher resolution models, defined by (5.2), might be more appropriate.

Table 5.4 presents the optimal CoMs, $M_A \cdot M_B$, at specific diurnal blocks BL_c and BL_d, that have the best performances (i.e. highest R^2 values). The three best models for each customer-class are shown, irrespective of whether these are given as products or rations of metrics (however the corresponding operations are shown in Table 5.4, in the 2nd column). For TR

(and I&C), the three best CoMs include Metrics-2, 5 and a single occurrence of Metric-1, i.e. normalised and z-score normalised mean values and normalised minimum values. In all three cases the diurnal-block pairs include Block-C, which indicates the importance of the mid-day demand-levels for distinguishing between domestic and non-domestic consumption, also discussed in the per half-hour analysis.

Table 5.4: Optimal CoMs for the diurnal-blocks analysis

Customer Sector	$M_A \cdot M_B$	BL_c	BL_d	R^2
TR and I&C	$M_2 \div M_1$	Block-C	Block-A	0.93
	$M_2 \div M_5$	Block-C	Block-A	0.92
	$M_5 \div M_5$	Block-C	Block-D	0.92
OR	$M_{11} \div M_4$	Block-A	Block-D	0.92
	$M_6 \div M_1$	Block-A	Block-A	0.90
	$M_4 \times M_5$	Block-D	Block-A	0.89
E7	$M_4 \div M_6$	Block-A	Block-A	0.75
	$M_{10} \times M_{12}$	Block-A	Block-D	0.74
	$M_1 \div M_4$	Block-C	Block-A	0.74

For the estimation of OR and E7 percentages, a more diverse set of metrics is shown, including Metric-1 (normalised mean), Metric-4 (z-score normalised maximum), Metric-6 (z-score normalised minimum) and to a lesser extent Metrics-11, 5, 10 and 12. These correspond, primarily, to Blocks-A and D, i.e. night and evening periods, which is, again, justifiable based on the assumption that E7 customers can be more accurately differentiated based on electricity demands during night-hours.

5.3.4 Customer-Class Disaggregation Results

The multiplicity of resulting estimators (i.e. successful CoMs) can be used to detect GSPs that can be considered as "outliers" of the analysis, such as GSPs with demands from the industrial sector (not I&C, but predominantly-industrial), for which the customer-class identification method has limited success. Examples of such GSPs are not included in the target-dataset and therefore the resulting models cannot account for their atypical consumption profiles, in terms of accurate percentage-contribution estimations (since the predefined customer-classes do not include a predominantly industrial-sector).

An example of the estimated percentages from different CoMs is shown in Figure 5.15 (a), for TR consumption. These results are based on the six metric-combinations with the highest R^2 values, i.e. the three best from the per half-hour analysis and the three best from the diurnal-blocks analysis. Marked in Figure 5.15 (a) is also the mean-absolute-deviation (MAD), given by:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)| \qquad (5.8)$$

where $n = 6$ for the six estimations and $m(X)$ is their mean value. MAD is therefore given as a quantification of the inconsistencies between the presented estimations.

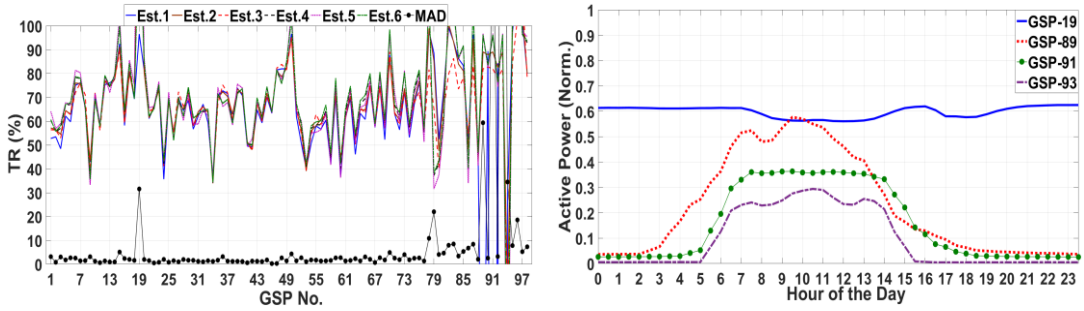


Figure 5.15: a) % of TR consumption for 98 GSPs based on the best 6 CoMs, 3 per half-hour and 3 per diurnal-blocks and b) GSPs with inconsistent results among the 6 estimations

Significant deviations from the mean (i.e. high MAD values) are shown for GSPs-19, 89, 91 and 93 which are above 30 % (and to a lesser extent for GSPs-78, 79, 94 and 96). The mean diurnal profiles of these four GSPs are shown in Figure 5.15 (b), in normalised values (3.3). These GSPs have also been identified in various instances in Chapter 3, due to their atypical demand patterns, in the diurnal and seasonal profiling analysis. GSPs-89, 91 and 93 are also the only three members of cluster-7 (C7), as presented in the clustering classification results in Section 5.2. They are in fact labelled as "factory-consumption" in the initial datasets, as provided from the corresponding DNO and can be considered as predominantly/exclusively industrial GSPs[18].

Apart from determining "outliers", the process of comparing the results from various estimations is used for the selection of the final set of CoMs from which the percentages for TR (and I&C), OR and E7 are calculated. Agreement between estimated percentages from various CoMs indicates that the same demand characteristics are determined even though different metrics are used. If the opposite is true, i.e. the resulting estimations show high levels of inconsistency, it implies that the various combinations are only successful at modelling the target percentages but fail when applied to the total number of GSPs. These results are presented in Figure 5.16, for TR (and I&C) in (a), for OR in (b) and for E7 in (c). Each plot shows the empirical cumulative distribution function (CDF) for the quantified inconsistencies

[18] They are not included in the target-dataset because the exact percentages from the industrial-sector are not known. Furthermore, demand patterns from the industrial-sector are assumed to be specific to each particular industry and generalised characteristics (similar to the morning/evening peaks for the residential-sector) are not available.

(in MAD values) among the 98 GPSs, when relying on the 5 best CoMs from the diurnal-blocks analysis, the 5 best CoMs from the per half-hour analysis and when considering both (i.e. for 10 CoMs, 5 from each method).

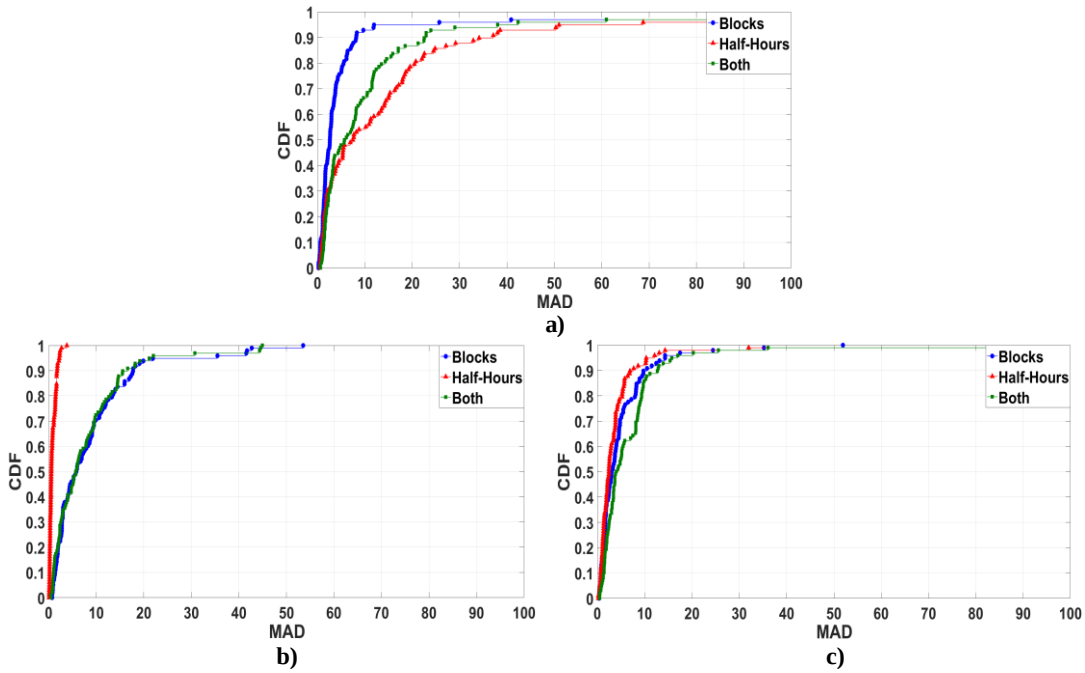


Figure 5.16: CDF for the mean-absolute-deviations in the CoM-estimations, based on the half-hour analysis, the diurnal-blocks analysis and both for a) TR (and I&C), b) OR and c) E7

For TR, plot (a), better consistency is shown for the blocks approach, with MAD below 10 % for more than 90 % of the sample, while only for ~50 % of the sample for the per half-hour analysis. For OR, very high consistency is shown for the per half-hour analysis, with 100 % of the sample below 10 % MAD, while the same performance is shown for ~70 % of the sample for the blocks analysis. Better consistency is also shown for the per half-hour analysis in the E7 case, plot (c), however both the per half-hour and the blocks analysis show MAD values below 10 % for ~90 % of the sample.

Therefore, the generalisation of the analysis, from individual half-hours to diurnal-blocks, produces resulting estimations of improved consistency for the TR customer-class and reduced consistency for the OR and E7 customer-classes. It should be noted that these results quantify the consistency of estimations but do not guarantee that the estimations are themselves accurate with respect to the actual consumption percentages. They are therefore measures of precision and not measures of accuracy. If systematic errors (biases) are included in the methodology, then it is possible that these will be included in a number of estimations. However, higher consistency shows that the resulting combinations-of-metrics are able to identify the same features. In all three cases (i.e. TR, OR and E7), more accurate estimations

with respect to the target-dataset are given from the per half-hour analysis, as indicated from the R^2 values in the previous section, as well as from the error between the target percentages and the estimated percentages, which is, on average, higher for the blocks approach than for the per half-hour approach.

Based on the above discussion, the final estimations for the percentage-contributions from TR (and I&C), OR and E7 consumptions are given according to the average values of: the 5 best CoMs from the diurnal-blocks analysis for the TR (and I&C); and the 5 best CoMs from the per half-hour analysis, for both the OR and E7. The resulting percentages for TR and E7, with respect to the first 6 clusters from Section 5.2 (excluding C-7 and single-GSP clusters-8, 9 ,10 and 11) are presented in Figure 5.17.

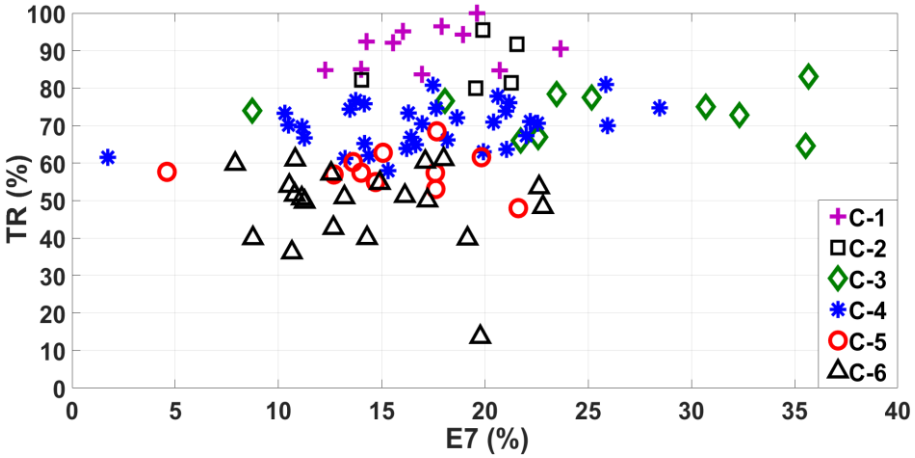


Figure 5.17: Percentages from total residential (TR) and economy-7 (E7) consumption

Clusters 1 to 6 are in the correct order from decreasing TR contributions to increasing I&C contributions, as in the predicted ordering discussed in Section 5.2. There is, however, a significant level of overlapping and the clusters are not clearly separable in the final results, particularly for TR consumption between 50 % to 80 %. A higher cut-off threshold level in the clustering procedure can produce a larger number of clusters with less overlapping, but this would also result in more single-GSP clusters with very specific consumption patterns.

The results are presented in more detail in Table 5.5, for all GSPs and their corresponding clusters from Section 5.2, excluding C7 and single-GSP clusters-8, 9 ,10 and 11. Table 5.5 also presents the mean TR percentages per cluster, as well as the standard deviations for GSP percentages within each cluster (mean values are shown under the individual cluster numbers and standard deviations are shown in brackets). Ordinary residential (OR) percentages are not presented, but similar to the I&C percentages, these can be calculated from the results according to (5.6) and (5.7).

Table 5.5: Estimated % to total consumption from TR and E7 customer-classes (nearest integer approximation)

Cluster	GSP No.	TR (%)	E7 (%)	Cluster	GSP No.	TR (%)	E7 (%)	Cluster	GSP No.	TR (%)	E7 (%)
C1 91(7)	***78***	92	16	*C4 70 (6)*	***4***	64	16	*C5 58 (5)*	***55***	62	20
	79	85	14		***5***	61	13		***56***	57	14
	82	95	16		***6***	77	14		***57***	63	15
	83	94	19		***7***	76	21		***67***	60	14
	84	85	21		***8***	67	22		***69***	55	15
	85	84	17		***10***	74	13		***72***	58	5
	87	92	14		***11***	58	15		***74***	57	13
	90	91	24		***12***	75	28		***75***	68	18
	92	85	12		***13***	76	14		***77***	57	18
	95	96	18		***14***	78	21		***86***	53	18
	97	100	20		***18***	70	11		***88***	48	22
C2 86 (7)	***15***	92	22		***20***	81	17	*C6 49 (11)*	***1***	54	23
	17	80	20		***21***	63	20		***2***	55	15
	47	81	21		***22***	67	11		***3***	50	11
	48	82	14		***23***	75	18		***9***	43	13
	50	95	20		***25***	71	20		***16***	60	8
C3 74 (6)	***59***	75	31		***27***	70	26		***24***	40	14
	61	65	36		***28***	65	14		***26***	54	11
	62	73	32		***29***	70	10		***30***	57	13
	64	67	23		***32***	67	16		***31***	61	11
	65	66	22		***33***	62	14		***34***	36	11
	66	78	23		***35***	71	23		***41***	51	13
	68	77	25		***36***	73	16		***42***	50	17
	76	77	18		***37***	71	22		***44***	60	17
	81	74	9		***38***	62	2		***51***	61	18
	98	83	36		***39***	73	10		***52***	51	16
					40	72	19		***53***	40	9
					43	66	18		***54***	52	11
					45	71	17		***58***	48	23
					46	64	21		***60***	40	19
					49	81	26		***63***	51	11
					71	65	17		***94***	14	20
					73	74	21				

While the limited number of GSPs with known percentage-contributions does not allow for direct validation, indirect validation can be offered based on previous analysis. The results are in agreement with the discussion provided in Chapter 3, Section 3.4, for GSPs with increased consumption during the weekends, compared with the general trend of increased consumption during the weekdays. It was hypothesised that these GSPs (approximately half of the Danish GSPs) are predominantly residential and that for the corresponding load composition, consumption is reduced for the days when people are at work and, therefore, spend less time at home. The current analysis shows that, indeed, these GSPs have some of the highest percentages of TR consumption and belong to cluster 1 (the corresponding correlation between estimated TR percentages and weekday/weekend demand differences has been presented in Chapter 3, Section 3.4). Higher TR consumption is also shown for the GSPs with increased seasonal variability compared to the daily and weekly variabilities, which was also assumed to be linked to the characteristics of residential consumption (Chapter 3, Section 3.3), as well

as for GSPs with stronger seasonal (per half-hour) correlations with temperature, as presented in Chapter 4, Section 4.6 (similarly, the resulting correlation between TR percentages and the *P-T* relationship have been presented in Section 4.6).

Furthermore, and according to (5.7), that the total percentage-contributions from the OR and E7 customer-classes sum up to the total residential percentages. This provides another form of validation since OR and E7 categories are estimated independently from the TR percentages and therefore high deviations with respect to (5.7) can indicate poor performance of the models.

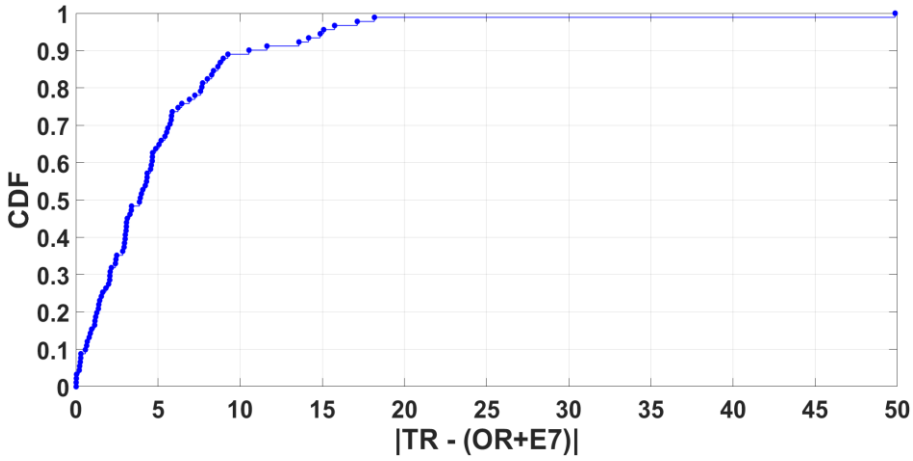


Figure 5.18: CDF of the % difference between OR+E7 and TR estimated % contributions

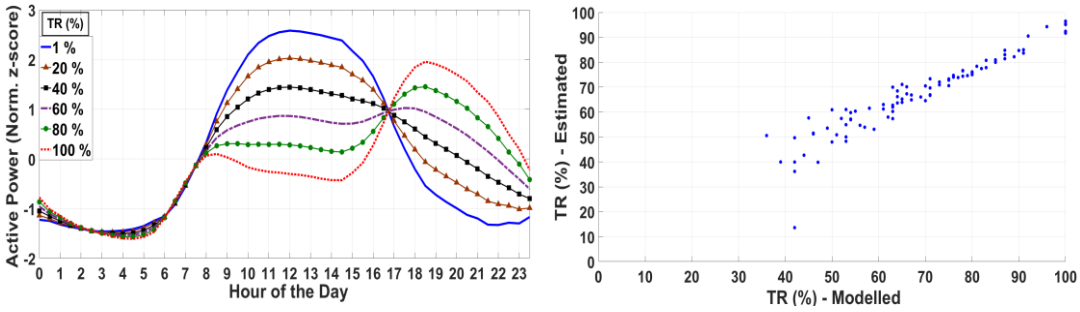
These results are presented in Figure 5.18 for all GSPs excluding clusters 7 to 11, as previously discussed. The error is below 5 % for approximately 65 % of the GSPs and below 10 % for approximately 90 % of the GSPs, indicating an overall satisfactory performance for the customer-class disaggregation.

5.4 Chapter Conclusions

Load composition, with respect to percentage contributions from different customer-classes, is in fact a determining factor of the characteristic diurnal consumption patterns (load profiles), as well as for the seasonal demand variability. The first is demonstrated by the ability of clustering algorithms and other feature-extraction procedures to group together electricity substations according to diurnal demand profiles. The second is shown by the inclusion of metrics such as the seasonal range of variations and the coefficient of determination (R^2 for seasonal *P-T* correlations) as identifiers of the corresponding percentages. However, and as demonstrated through the comparison of the clustering-classification (Section 5.2) and the more detailed customer-class disaggregation (Section 5.3), care should be taken when relying

solely on clustering for GSP-classification. These clusters show intersections and a, relatively, wide range of variability for the decomposed percentages (Figure 5.17 and Table 5.5), which indicates that accurate estimations require the identification of more fine/detailed differences among the considered GSPs. These patterns, as presented in Section 5.3 and particularly in the per half-hour analysis, are in good agreement with theoretical/expected markers of load composition, such as: demand levels during the mid-day period (associated with commercial consumption), morning and evening peaks (associated with residential consumption) and the seasonal range of demands during evening and night hours (associated with economy-7 residential demands).

It is assumed that the analysis is more accurate for the Scottish GSPs, due to the fact that the target (training) dataset was comprised exclusively of Scottish consumption statistics. This is supported by the fact that a relatively higher level of inconsistencies has been shown for the remaining number of GSPs, in Figure 5.15 (a). However, the methodology presented was sufficient to identify GSPs that largely deviate from common demand characteristics and it is expected that with a larger training dataset the approach can be improved and extended in order to account for a larger set of customer-classes. In the same context, OR and E7 customer-classes are expected to contain more inaccuracies than the corresponding TR percentages, since these groups have characteristics more specific to the Scottish dataset.



a) b)

Figure 5.19: a) modelled diurnal profiles for various TR contributions and b) correlation of estimated and modelled TR percentages

The relationships established between the various metrics (Table 5.1) and the percentages from different customer-classes, can be used for the development of models, able to generate diurnal (or seasonal) demand profiles according to selected percentages, in order to be used as inputs for subsequent studies (and in cases where actual data is limited or unavailable). Initial results from this approach are shown in Figure 5.19 (a). The 6 curves are based on the resulting estimations for total residential demands, which are used to produce diurnal demand profiles, for specified percentages from the TR customer-class. Figure 5.19 (b) shows the relationship (and strong correlation) between the TR percentages for 91 GSPs, as estimated from Section

5.3, and the corresponding TR percentages from the modelled diurnal patterns. Modelled diurnal profiles are generated for a range of TR [0 100] in steps of 1 %. Then, each profile is correlated with the actual mean diurnal profiles of individual GSPs. The TR-level at which the correlations are maximized is what is shown as the x-axis in Figure 5.19 (b). The approach can be expanded to include the seasonal components, as well as metrics defining the range of variations and therefore be able to reconstruct demands according to a more diverse set of consumption characteristic.

Classification and disaggregation of demand measurement according to end-user sector and consumption characteristics is important for network operators as well as for electricity suppliers. Possible applications include more informed tariff formulations, inputs for the development and evaluation of demand-side interventions, assessment for proposed DG-integration and planning considerations. Customer-class disaggregation based on the analysis of direct consumption/demand datasets allows for the development of models that are more specific to locations, DNOs and the corresponding customer mixtures, with less reliance on survey-based data or the necessity of installing measuring devices at the LV-level (individual households, businesses, industries, etc.).

Chapter 6: Load Disaggregation

This chapter presents methodologies for the disaggregation of total active power demands into generic and specific load components. In the first instance, these are defined according to their contributions to the total demand variability whereas, in the second case, two approaches are presented for the disaggregation of demands into heating, cooling, lighting and seasonally variable non-thermal loads. The results are demonstrated on selected GSPs, that have been shown to have distinctively different load compositions, according to the customer-class disaggregation approach discussed in the previous chapter, and the final estimations, which correspond to the seasonal percentage contributions, are presented based on percentile values for a total number of 77 GPS.

Section 6.1 presents the decomposition of loads according to the seasonal and diurnal variabilities. Annual and daily base loads are discussed in Section 6.1.1, which are the minimum recorded demands per GSP, throughout the duration of one calendar year and for each day of the year separately. Base loads are also determined at each diurnal period (i.e. half-hour) in Section 6.1.2, using three different approaches and the results are then utilised for the estimation of base temperature values (i.e. threshold or balance-point temperatures), which mark thermal-comfort levels and are therefore used in the subsequent sections for disaggregation. Similarly, Section 6.1.4 presents two different approaches for the estimation of solar irradiance (and solar elevation angle) base levels. The temperature and solar bases are specific to the weather-demand interaction for the UK, as only the English and Scottish datasets have been used. In Section 6.2, the diurnal, seasonal and seasonal per half-hour base active power estimations are used for the decomposition of demands into four different portions, which are then expressed in terms of base, intermediate and peak load contributions.

Section 6.3 introduces three multiple regression models, using active power as the dependent variable and various combinations of reactive power, temperature and solar irradiance/elevation angle as the sets of independent variables. These are accordingly adjusted for the disaggregation of different load types and the results are presented in Sections 6.3.1, 6.3.2 and 6.3.3, for thermal heating, thermal cooling and lighting loads, respectively. In Section 6.4, a novel approach is presented, for the disaggregation of total demand into seasonally variable loads, corresponding to thermally-related and non-thermally related demands, based on the active power reactive power relationship and relying on power-factor data transformations.

6.1 Constant and Variable Loads

An intuitive and logical starting point of the load disaggregation procedure is making the distinctions between the variable and constant (base) portions of the aggregated demand envelopes. In this context, base loads can be defined as the minimum required active/reactive power demands and can be calculated for individual GSPs in a number of ways, depending on the selected time-frame, target application or indented use for further analysis. Section 6.1.1 presents the annual and daily base loads, while the per half-hour base and variable portions of the load are further decomposed with respect to the diurnal and seasonal cycles, as presented in Section 6.1.2. The results are combined and used to estimate base (i.e. threshold or balance-point) temperature levels, presented in Section 6.1.3, while Section 6.1.4 presents the corresponding solar irradiance and elevation angle bases. Specific definitions, estimation approaches and discussions are provided in the subsequent sections.

6.1.1 Annual and Daily Base Loads

The annual base load refers to the overall minimum measured demand from a particular GSP within the period of one calendar year. An example of the annual base load is shown in Figure 6.1 (a), in the diurnal perspective, i.e. all daily profiles from a one-year period and in (b), in a seasonal perspective, i.e. all half-hourly profiles throughout a one-year period. The annual base is calculated as the minimum recorded demand or, alternatively, as the 1st or otherwise percentile of recorded demands (indicated in Figures 6.1 (a) and (b) as the black-dotted and black-dashed lines, respectively).

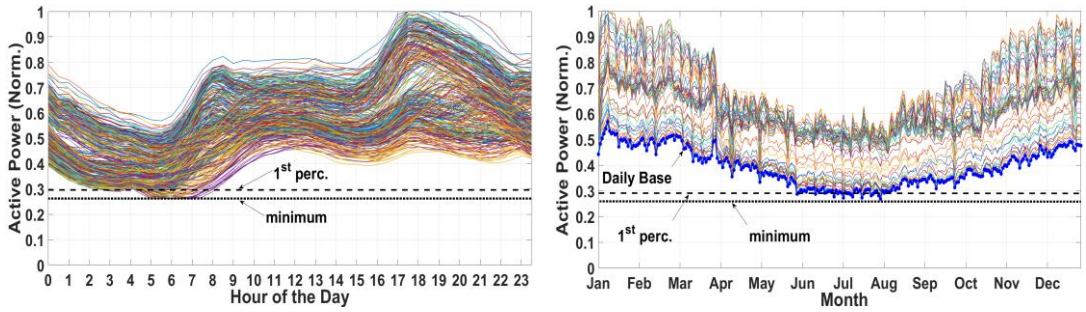


Figure 6.1: Examples of a) the annual base load in daily profiles and b) the annual base load and the daily base loads in half-hourly seasonal profiles

Large deviations between the minimum and percentile-minimum values can be indicative of outliers in the measured demands (e.g. zero values) and the difference between the two metrics can be used to identify them. When no outliers are present the final estimated annual base can be expressed as the average of the two. Confidence intervals and error margins can also be calculated, based on the two estimates but, generally, these tend to converge within narrow

limits, i.e. annual base loads are clearly evident from the demand profiles and are well defined. Annual base values have also been discussed in Chapter 3, Section 3.7, in the context of demand profiling and for determining the probability of occurrence of maximum and minimum demands within different time-frames. In Figure 6.1 (b), the daily-base loads are also shown (blue-solid line). These are the minimum required loads for each day of the year, also discussed in Chapter 4 Section 4.5, as the minimum-daily demands used in the correlation analysis. The minimum of the daily-base loads, in a one-year period, is equivalent to the annual base, as shown by the intersection of the blue-solid and black-dashed (and dotted) lines is Figure 6.1 (b).

Figure 6.2 (a) shows the resulting active and reactive power annual bases for 77 GSPs, normalised with respect to peak demands per GSP (3.3), and Figure 6.2 (b) shows the corresponding bases in terms of actual (non-normalised) demand values. No significant correlation can be reported in the first case; however, the correlations can be improved by using the actual values, as shown in (b).

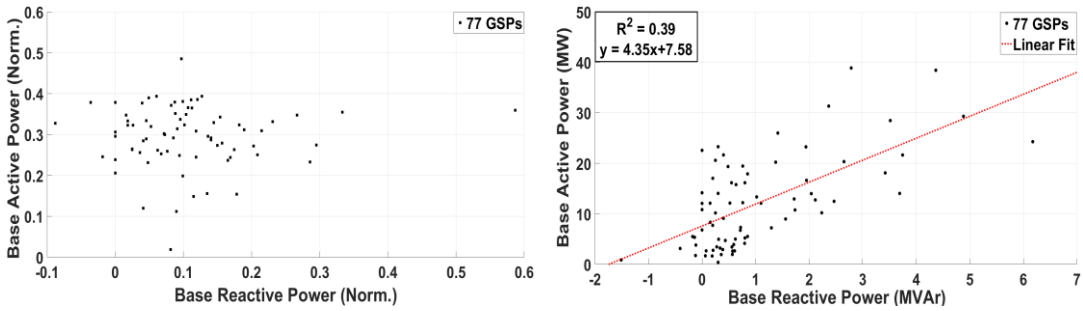


Figure 6.2: Relationship between annual active power base and annual reactive power base for a) normalised values and b) actual values

The differences in the distributions of data-points between (a) and (b) are due to the relative "distances" between demands and peak demands connected to issues of normalisation, as discussed in Chapter 3, Section 3.2. When the effects of normalisation are removed, the correlations are improved, but still below $R^2 = 0.4$, as shown in (b). While there is a general tendency of increasing base reactive power for increasing base active power, the relationship shows exceptions and high variability, i.e. between base reactive powers of $0 \leq MVAr \leq 1$, there are base active powers of $0 \leq MW \leq 20$. The negative reactive power base values are most probably associated with reactive power flows from distributed/renewable generation and/or an increase in capacitive loads, at the corresponding GSPs.

Figure 6.3 (a) shows the relationship between mean active power and base active power and Figure 6.3 (b) the relationship between mean reactive power and base reactive power, for 77 GSPs, in both cases. Actual (non-normalised) values are presented, following the discussion

provided for Figure 6.2. The results show strong correlations between mean active power and base active power (R^2 ~ 0.9) and moderate/strong correlations between mean reactive power and base reactive power (R^2 ~ 0.65). This demonstrates that, at least in the case of active power, the annual bases can be estimated with high accuracy based on mean demands (and vice-versa), following the equations provided in Figures 6.3 and based on the analysis of the available datasets.

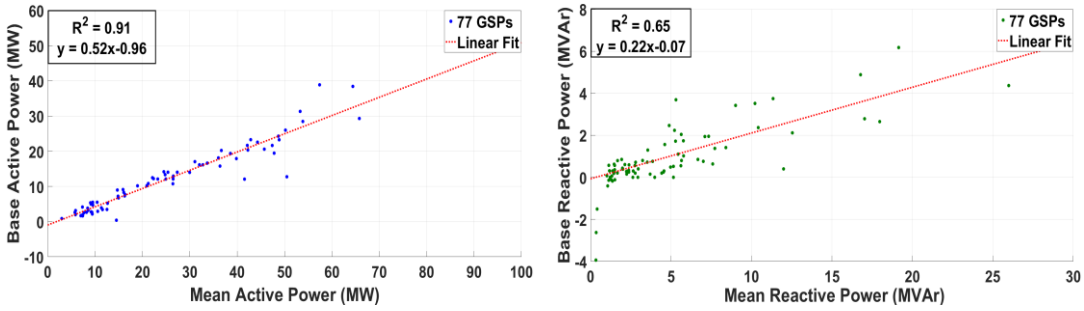


Figure 6.3: Relationship between: a) mean active power and annual base active power and b) mean reactive power and annual base reactive power

Aggregate base-loads or system (grid) base-loads are of interest for electricity generation and demand/supply balancing, as they are used to define the minimum power required to be delivered by the system. In the same context, distinctions are made for the intermediate loads and peak loads, in terms of modes of generation and dispatch periods. For the purposes of load disaggregation, base demand is useful because it includes loads that remain relatively constant for the whole year and irrespective of the hour of the day (Figure 6.1), such as stand-by consumption from electronic devices, wet loads and cold appliances (although some seasonal variability exists), etc. In reality however and for aggregated demands, such as from MV-GSPs, the base includes a larger variety of loads that remain relatively constant due to their aggregation, i.e. it cannot be assumed that the base loads are homogeneous with respect to load type nor that variability is absent from the individual load types. This can be seen from Figure 6.2 (a) where base loads are within 0.2 to 0.4 (in per unit) of peak demand and from Figure 6.3 (a) that shows a beta-coefficient of ~0.5 for the base-to-mean relationship. Base-loads representative of individual household consumption, such as from data acquired from smart-metering devices, shows that the ratio of base to peak/mean demand is lower than the one shown from this analysis, as it is possible for individual households to reach levels of very low power consumption (e.g. during night hours)[19]. Due to these high ratios with respect to

[19] Analysis has also shown a moderate negative correlation ($r \sim -0.4$) between the ratio of base-to-mean and the estimated percentage-contributions from total residential demands (TR), from Chapter 5. Predominantly residential GSPs (i.e. TR > 90 %) have base-to-mean ratios between 0.3 and 04.

peak/mean demands, base loads in MV-GSPs can define the overall minimum requirements, however the constituent load categories that make-up base loads are not necessarily stationary nor is their range necessarily restricted to the base levels. This is also supported by Figure 6.1 (b), which shows the daily-base load for each day throughout one calendar year, were distinctive seasonality is evident. The overall constant levels of annual base loads (per GSP) present a particular problem for disaggregation because the absence of apparent variability (of active and reactive power) limits the suitable approaches that can be used to disaggregate this portion of measured demand.

6.1.2 Base Loads Per Diurnal Periods

The base loads presented in Section 6.1.1, both annual and daily, make no distinctions between the different levels of demands within the diurnal period. An extension of the analysis therefore includes determining base loads at each half-hour of the day, or according to available sub-daily data resolution. The benefit is threefold:

1) per half-hour base loads can be used to determine the seasonally-constant portion of the load for each diurnal period, which is by definition higher than the annual base and the difference of the two shows the portion of demand that is constant throughout the year but varies from one half-hour to the next. Annual and per half-hour base loads are equal only at the half-hour at which annual base is found, as shown in Section 6.1.1, Figure 6.1 (a).

2) for each half-hour, the difference between the half-hour base load and the half-hour peak load, inherently includes the seasonally variable portion of the load, which can be associated with demand variations of particular load types (expected) to be used at specific period(s) of the day.

3) the results can be used to determine the base/threshold temperature values, which can be defined as the levels at which any further increase/decrease in temperatures results in an increase of overall active power demand, due to (but not exclusively) cooling/heating loads being switched on (although there exists a latency temperature range, i.e. conformable temperatures, and therefore bases can be defined separately for cooling and heating loads).

Accordingly, three different approaches are considered and combined to give the final estimations for the quantities discussed above.

Base Load/Temperature Estimations – Method 1: The first method involves a 3rd degree polynomial best fit, for active power and temperature, for each GSP, at each half-hour of the day and using weekdays only. The regression approach is similar to that discussed in Chapter

4, Section 4.3, but includes the additional x^3 and x^2 terms with their corresponding β -coefficients, for x – temperature and $f(x)$ – estimated active power:

$$f(x) = \beta_1 x^3 + \beta_2 x^2 + \beta_3 x + \beta_0 \quad (6.1)$$

The regression model is also adjusted to minimize the sum of absolute residuals (LAR model), in contrast with the conventional least squares method (OLS model) which minimizes the sum of squared residuals. The result is a robust version of regression that assigns less weights to the data-outliers [214]. An example of the analysis is shown in Figure 6.4, for GSP-14, at 11:00 hours.

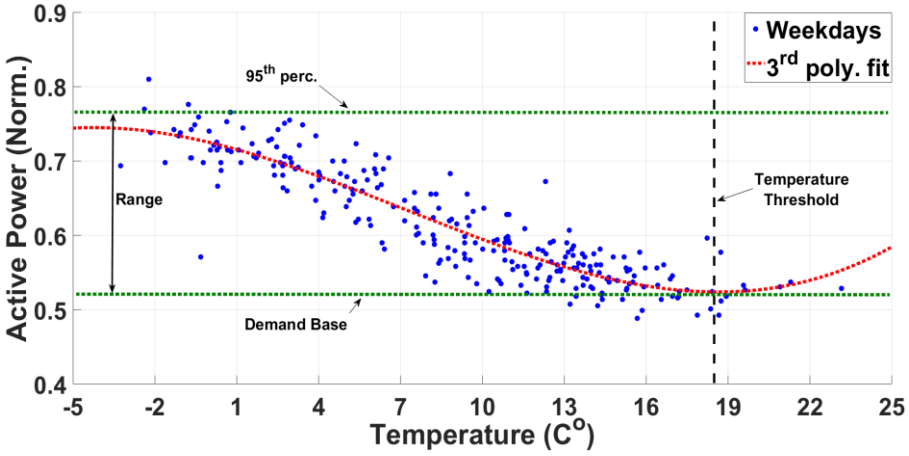


Figure 6.4: An example of the estimation of per half-hour base active power, seasonal range and temperature threshold, for GSP-14 at 11:00 hours (weekdays only) – Methods 1

When the LAR regression is computed, the best fit function (6.1) is retrieved and used to estimate the corresponding base demand and temperature threshold values such that:

$$P_B(t) = \min_{a_t \le x \le b_t} f(x)_t \quad (6.2)$$

$$T_{thre.}(t) = \underset{a_t \le x \le b_t}{argmin}\, f(x)_t \quad (6.3)$$

where $P_B(t)$ and $T_{thre.}(t)$ are the base demand and temperature threshold values at half-hour t, $f(x)_t$ is the polynomial best fit as calculated at each half-hour using (6.1) and $a_t \le x \le b_t$ is the range of temperature values at each half-hour throughout the year. The range of seasonal variations, i.e. $P_{SR}(t)$, is calculated as the difference between the base demand and the 95th percentile (maximum) demand, as shown in Figure 6.4, i.e.:

$$P_{SR}(t) = P_{95th}(t) - P_B(t) \quad (6.4)$$

This approach performs well in cases where there is a moderate to strong correlation between active power and temperature and particularly when the relationship is non-linear, such as in

the presence of significant levels of both heating and cooling loads. The geographical locations corresponding to the available datasets are characterised by cold/temperate climates and therefore do not have cooling demands comparable to heating demands, however the desired "turning points", at the temperature threshold values, can be correctly estimated for approximately 75 % of the total number of analysed GSPs. In the cases where this method fails, the errors are identifiable from the resulting active power and temperature threshold estimations being "unrealistic", i.e. outside the excepted range of threshold values. To account for these cases and to automate the selection process, a constraint is set so that the active power base must correspond to the first derivative of the polynomial fit being equal to zero, such that (6.2) and (6.3) are valid only when:

$$f(x_i)'_t = 0 \;\; and \;\; \min_{a_t \le x \le b_t} f(x)_t = f(x_j)_t$$

$$subject\ to: i = j \qquad (6.5)$$

Base Load/Temperature Estimations – Methods 2 & 3: The second and third methods make no use of the active power-temperature relationships and instead rely on the seasonal, per half-hour profiles of active power demands, using actual and smoothed values as shown in Figure 6.5, for the same GSP and at the same time, i.e. GSP-14 at 11:00 hours.

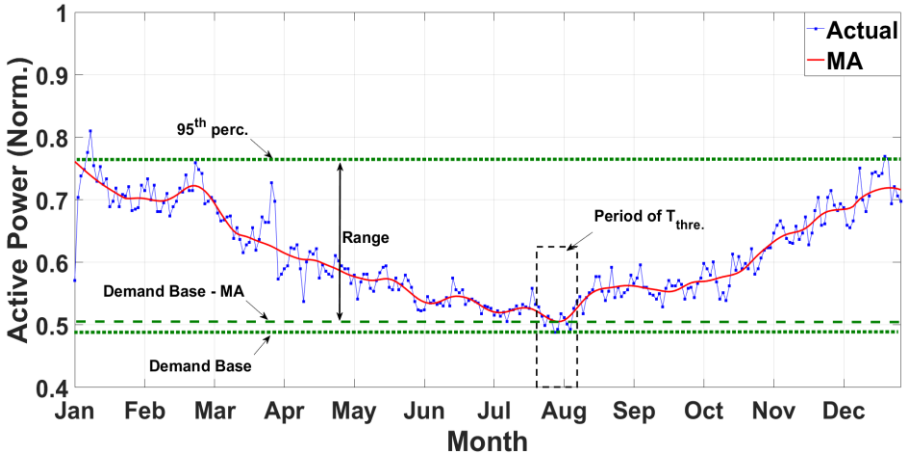


Figure 6.5: An example of the estimation of per half-hour base active power, seasonal range and temperature threshold for GSP-14 at 11:00 hours (weekdays only) – Methods 2&3

Base active power demands for each half-hour of the day are calculated from the minimum demands; in the first case using the actual normalised demand values (Method-2) and in the second case using the "smoothed" demand values, as calculated using a moving-average filter, given by (3.13), of window length ±20 weekdays (Method-3). The range is defined as the difference between any of the two base estimations and the 95th percentile maximum values, as in (6.4). For the majority of the GSPs, the two base estimations are at, approximately, the

same levels. However, the use of filtered active power values produces base estimates that are less sensitive to data outliers, similar to the percentile base values used in Section 6.1.1. Marked on Figure 6.5 is also the period from which the corresponding temperature values are selected, in order to estimate the per half-hour temperature thresholds, i.e. average value of the temperatures within the window defined by the ten lowest active power demands.

6.1.3 Base Temperatures

Base (i.e. threshold or balance-point) temperature values have been used in previous studies, particularly in the context of estimating heating and cooling degree days (HDD & CDD), as in [215] and [216], which can then be used for estimating demands for heating and cooling loads. A range of base temperatures is typically chosen from which the corresponding HDD and CDD are calculated, at each particular base. In other studies, the bases have been calculated from the demand-temperature relationships, with approaches similar to the ones used in this thesis (e.g. in [217] for Spain, but not up to a half-hourly resolution). There is no single universally agreed temperature threshold value and the choice depends on various factors including: the effects of other meteorological conditions such as relative humidity (as discussed in Chapter 4), the efficiency of a building's insulation (for more specific studies) and the familiarization (of customers) with particular weather conditions, which also relates to the geographical location of interest. In the UK, base temperatures are usually defined at around 15.5 C^o for HDD and around 22 C^o for CDD [218], although these can vary according to each specific study.

The approaches presented in Section 6.1.2 can be used to provide estimates for base temperatures that are derived directly from demand measurements and from demand-temperature relationships that are tailored to specific locations, populations, building types, perceived comfortable temperature levels, etc. Unlike conventionally used base temperatures, these estimates are also allowed to vary within the diurnal cycle, accounting in this way for changes with respect to the daily periods. Figure 6.6 shows the results for the estimated temperature threshold values, based on analysis conducted for the UK-GSPs[20] and for the three methods discussed in Section 6.1.2.

For all three methods, there is a distinctive daily pattern with lower base temperatures during night and early morning hours that is gradually increasing through the day, reaching a peak at

[20] Only the UK GSPs are used in order to give more climate/location specific estimations for the threshold temperature values and also because these are the datasets that are used subsequently for thermal heating/cooling load disaggregation. As discussed in Chapter 3, the UK data corresponds to the North of England and South/Central Scotland.

around 16:00 to 19:00 hours and then decreasing again. As the threshold is increasing through the day until the evening hours, this implies that heating equipment is turned on at higher temperatures than during the night or early morning. If the resulting base temperatures are indeed representative of perceived comfort levels, the diurnal patterns could be a result of people's daily schedules, in a sense that during periods when people are active at home, their response is also more sensitive to decreasing outdoor temperatures.

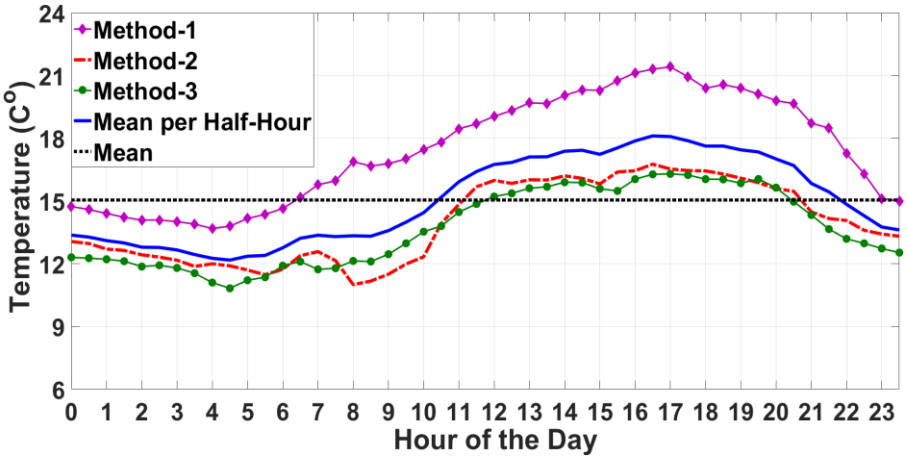


Figure 6.6: Estimated per half-hour and mean base (threshold) temperatures

Concerning the relative "heights" of the curves; method-1 is probably an overestimation of the threshold values, resulting from the application of a 3rd degree polynomial fit (Figure 6.4). Although it more closely matches the 15.5 C° mentioned in literature during the night and early morning hours (than the other methods during the same period), the values increase up to 21 C° for afternoon and evening hours, which can be considered unrealistic, at least for heating base temperatures (the values are closer to threshold limits that correspond to demand for cooling). Methods-2&3 produce similar results, indicating that using the minimum values of either actual or smoothed active power has no significant impact on the temperature threshold estimations, since in most cases both of these sets of minimum values correspond to the same yearly periods (as shown in Figure 6.5). Marked on Figure 6.6 is also the curve for the per half-hour mean values, as calculated from all three methods, ranging from ~ 12 C° during early morning hours (05:30) to ~18 C° during afternoon hours (17:00), as well as the "global mean", calculated as a single value from all half-hourly estimations, of all three methods. This is at a level of 15.05 C°.

Since the study of relative comfort temperatures includes psychological, physiological and behavioural considerations, the validation process is not straight-forward. The only conclusion that can be confidently stated is that the results are based on the analysis of actual demand responses and are therefore representative of the recorded changes in demands at MV GSPs.

The results are also in agreement with heating base values mentioned in literature. A possible systemic error in the presented analysis relates to the fact that for GSPs corresponding to customer/households of higher percentages of gas-based heating systems (or other technologies, including DG-renewable energy systems), the active power responses are compromised and can be less representative of the actual electric heating load responses. It should also be noted that although no distinctions are made between heating and cooling base temperatures, the range of the results shown in Figure 6.6 indicates that they both have effects on the estimations, primarily regarding Method-1, which reaches levels of up to ~21 C^o.

6.1.4 Base Solar Irradiance and Base Elevation Angle

Estimating ambient solar irradiance levels which correspond to the commencement of artificial lighting (i.e. periods at which lights are switched on/off), directly from demand/solar irradiance measurements, is more complicated than the estimation of temperature threshold levels. As demonstrated in Chapter 4, solar irradiance is not as strongly correlated with active power as temperature is, while there are strong dependencies between the two explanatory variables, particularly when the solar irradiance levels are replaced and represented by the "smooth" continues measurements of the solar elevation angle. Furthermore, and related to the correlation levels, there is a noticeable "inertia" with respect to people's responses when it comes to adjusting artificial lighting due to changes in ambient light, e.g. an increase in solar irradiance due to changes in cloud coverage is not immediately followed by switching-off artificial lighting, at least in the absence of automatically controlled lighting equipment. Although this assumption is generally and primarily based on personal experience, if the reverse was true, it would have probably been evident by stronger correlations between active power and solar irradiance (particularly when the effects of temperature are controlled for, as it is shown in Figure 6.8). These limitations are better illustrated in Figure 6.7, following a similar approach for the determination of base solar irradiance levels as the one presented for temperature (Figure 6.4), but this time for 17:00 hours.

While the polynomial best fit is successful in capturing the base active power demand, the solar irradiance base levels range from ~0 to ~0.7, in normalised values (x-axis) and therefore there is no clear "turning point" solar irradiance, similar to the one that has been shown for temperature. Furthermore, and in contrast with the base temperature determination, a 2^{nd} degree polynomial fit is applied here. This choice is based on a comparison between various fitting functions, from which the selected one demonstrated better consistency over the 48 diurnal periods and available GSPs, in a sense that it better defines minimum turning points that are in agreement with constraints similar to (6.5).

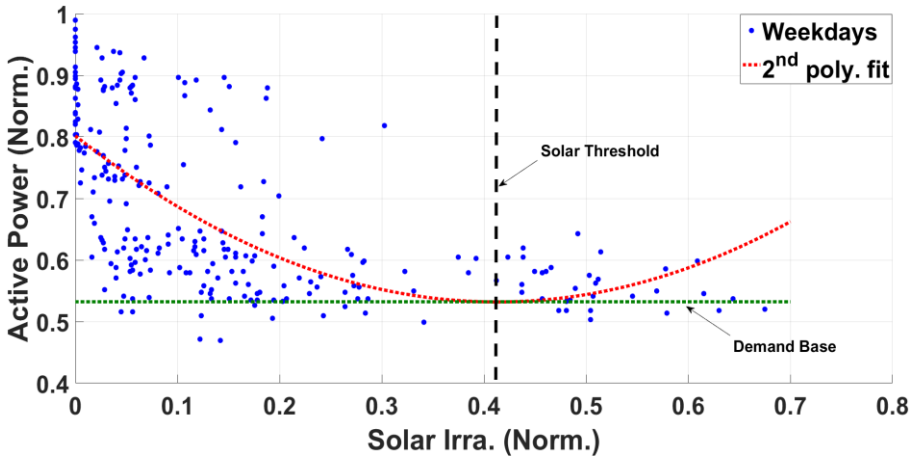


Figure 6.7: An example of the estimation of per half-hour base (threshold) solar irradiance for GSP-14 at 17:00 hours (weekdays only)

As it is shown in Figure 6.7, the resulting base (solar threshold) is an overestimation, since more noticeable increases in active power demands are shown for lower solar irradiance values, i.e. between ~0.2 and ~0.3, where it would be more appropriate to select the base.

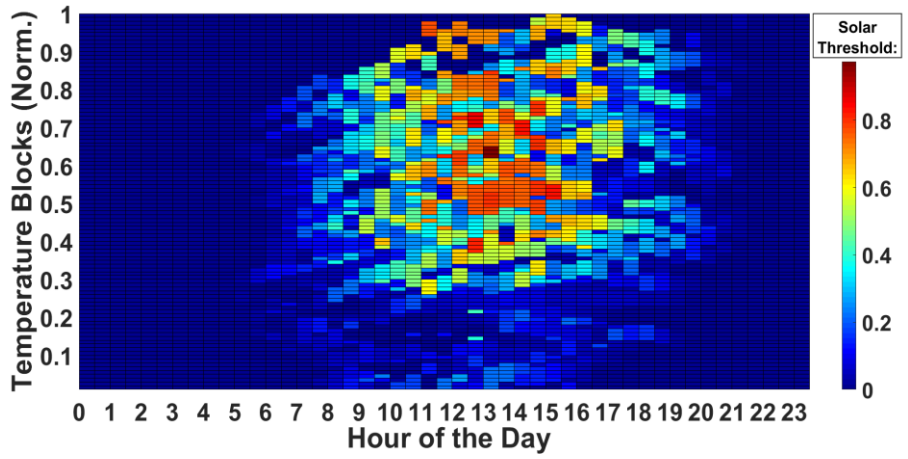


Figure 6.8: An example of the estimation of per half-hour base (threshold) solar irradiance at constant temperature levels, for GSP-14 (weekdays only)

A second approach is considered in which case the base solar irradiance levels are estimated at constant temperatures, in an attempt to mitigate the effects of the dependencies between the two explanatory variables. Since active power demand is primarily determined by temperature (in the seasonal perspective and for the majority of available GSPs), controlling for temperature differences by constraining the temperature levels within narrow limits (i.e. blocks) of ± 0.05 per unit, produces a wider range of base solar irradiance estimates. An example of the results is presented in Figure 6.8 for GSP-14, showing a tendency of increasing solar threshold values for increasing temperatures, particularly between 0 to 0.5 per unit, with peaks coinciding with the majority of the temperature data-points, i.e. within 0.5 and 0.8.

The final estimations are based on both methods (polynomial and blocks-analysis) and for solar irradiance as well as for solar elevation angles, thus giving a total of four diurnal profiles. Both variables are normalised with respect to the maximum overall measurement in each dataset (per-unit), to allow for comparisons and to produce final base values that are applicable for further analysis, irrespectively of which of the two is selected as the input variable. The results are presented in Figure 6.9.

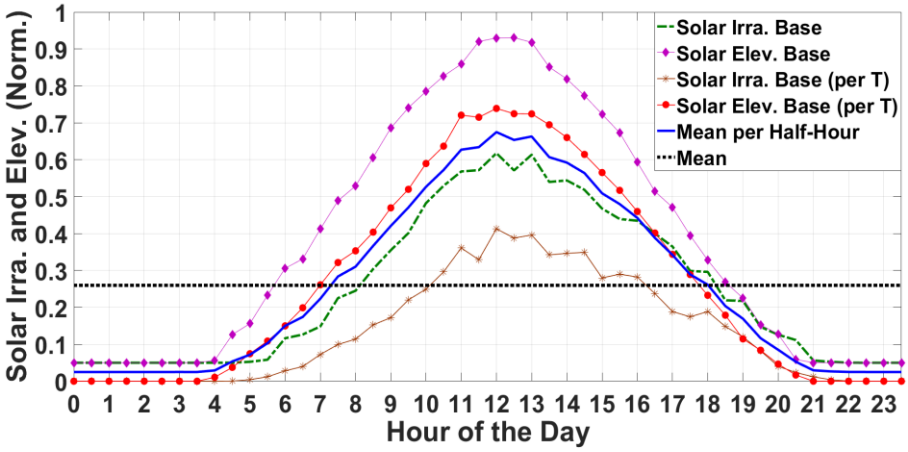


Figure 6.9: Resulting estimations for base (threshold) solar irradiance and elevation angles

While the controlled temperature-blocks approach (Figure 6.8) produces estimates that are significantly lower than the corresponding estimates from the polynomial method (Figure 6.7), in both cases the characteristic diurnal pattern results in values that can be considered overestimations of solar thresholds (this includes the mean per half-hour results from all four estimations, shown as the blue-solid line in Figure 6.9). A more reasonable base value is calculated by taking the average of all four estimations and over all half-hours of the day. This is represented by the black-dotted line in Figure 6.9, which is at a level of ~0.27 per unit.

These results are of course rough estimations and based on some necessary simplifications. In particular, although solar irradiance and solar elevation angles are primarily determined by the same underlying process, they are neither equivalent nor perfectly correlated, as solar irradiance is affected by atmospheric phenomena (e.g. cloud coverage) and includes components of direct, ground-reflected and diffused irradiation. Furthermore, these measurements are unavailable (by definition) for night-hours, and the diurnal patterns in Figure 6.10 cannot be considered representative of the sunlight-artificial lighting relationship, in the same way the patterns in Figure 6.6 (i.e. temperature thresholds) can. Nevertheless, and due to the absence of solar irradiance threshold values from literature, the results are used in the subsequent lighting-load disaggregation procedures.

6.2 ''Naïve'' Disaggregation

The results presented in the previous sections can be used to summarise distinctions among loads, as shown in the example in Figure 6.10, concerning weekdays only, for GSP-14. This form of disaggregation does not separate total demand into particular load-types, but it is rather an extension of profiling that aims to decompose total demand into constant and variable portions, according to the estimated bases and the extent of seasonal load variability.

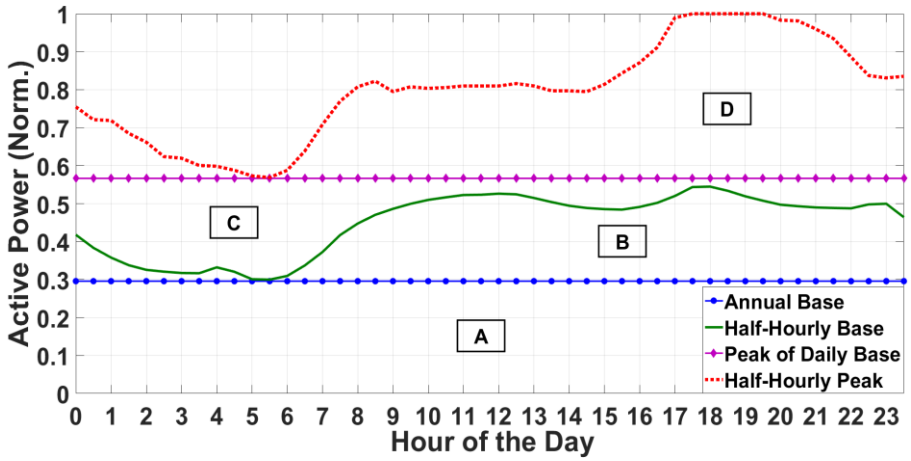


Figure 6.10: Resulting load distinctions in the diurnal perspective, for GSP-14

Marked on Figure 6.10 are the annual base and the maximum of the daily-bases, as discussed in Section 6.1.1 (Figure 6.1), as well as the half-hourly base values and the half-hourly peak values, from Section 6.1.2. The portion (area) of the load marked as A corresponds to the annual base, i.e. the minimum required and constant active power demand throughout the year. Area-B is the difference between the annual base and the individual per half-hour bases. The aggregated loads in Area-B are therefore unchanged with respect to the seasonal variations, but variability exists within the diurnal cycle. Area-C is the difference between the maximum of the daily-bases and the half-hourly bases and Area-D is the portion of demands with seasonal variability, as defined by the loads above the maximum of the daily-bases.

Therefore, the total of Areas A and B corresponds to demands up to the minimum measured active power at each half-hour of the day, i.e. a characteristic day which has half-hourly demands that are the minimum recorded throughout the year. The total of Areas B and C corresponds to the seasonal range of minimum daily demands and the total of Areas C and D corresponds to the range of seasonal variations, above the minimum recorded demands at each half-hour of the day.

These distinctions are illustrated in Figure 6.11, with respect to four different GSPs, i.e. GSPs-14, 15, 3 and 52. According to the customer-class disaggregation results, presented in Chapter

5 Section 5.3, GSPs-14 and 15 can be characterised as primarily residential (approximately 80 % and 90 % TR consumption, respectively), while GSPs-3 and 52 are considered mixture buses with higher percentages of commercial and industrial loads (approximately 50 % I&C consumption each). The y-axis in Figure 6.11 is the percentage of the corresponding areas with respect to the area defined by the maximum per half-hour demands (i.e. peak demands throughout the year, red-dotted curve in Figure 6.10).

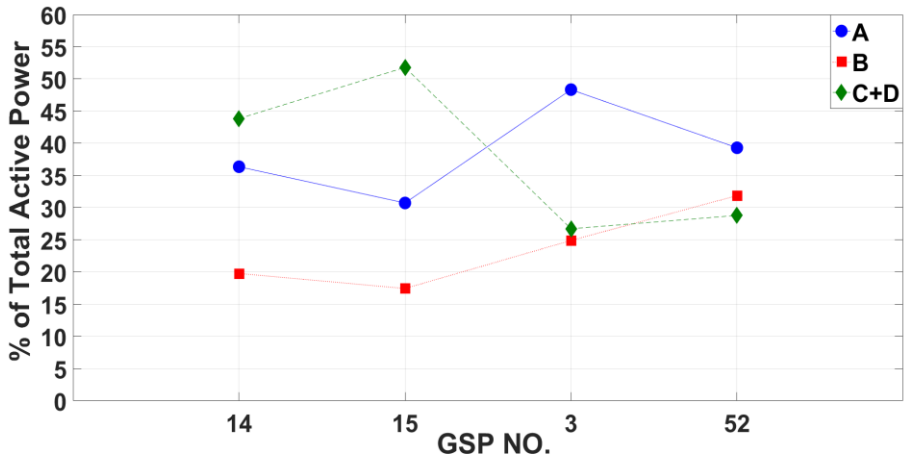


Figure 6.11: Resulting load distinctions for GSPs-14, 15, 3 and 52, for active power

Figure 6.11 shows that the residential GSPs have a higher percentage of loads corresponding to the sum of Areas C and D, i.e. the total seasonally variable portion of the load, which can be attributed (at least partly) to higher contributions from electric thermal loads, compared to the commercial GSPs. This conclusion is further supported by the estimated percentages of economy-7 (E7) residential consumption (Chapter 5, Section 5.3), which shows that residential GSPs-14 and 15 have approximately 20 % E7 contributions, compared to approximately 10 % and 15 % for GSPs-3 and 52.

It should be noted however, that these distinctions are not strong predictors of the relative contributions from the various customer-classes, which requires a more detailed pattern identification, as discussed in Chapter 5. Nevertheless, there are general tendencies, for example: increased total residential contributions (TR) are associated with increasing percentages (to total demand) from Areas C and D, i.e. seasonal variability correlates with domestic consumption; and decreasing contributions from TR are associated with increasing annual and half-hourly bases, i.e. Areas A and B. This is an indication of a, relatively, more constant load composition from the non-domestic sector. These results are supported by the analysis presented in Chapter 3, Section 3.3 (Figure 3.6 – Fourier analysis), which showed that, for the corresponding GSPs (14, 15, 3 and 52), the diurnal modes of variability had more

"weight" for the commercial GSPs, while the seasonal variability was more important for the residential GSPs.

The extension of the analysis presented here was therefore developed to allow for the decomposition of the variable and non-variable portions of active power demand and for expressing them as percentages to total demand. The approach can be used to determine the base, intermediate and peak portions of active power, for individual GSPs, or for groups of GSPs of the same network and depending on the selected definitions for these three load categories. For example, the peak demand, for each day of the year, is always included in Area-D and, similarly, daily-base loads are always included in Areas A, B and C, while the absolute minimum requirements are determined by Area-A. The cumulative distribution function (CDF) for demands measured at GSP-14, are presented in Figure 6.12, together with the four areas. In this context, Area-A can be considered the base load, Areas-B and C as the intermediate load and Area-D as the peak load.

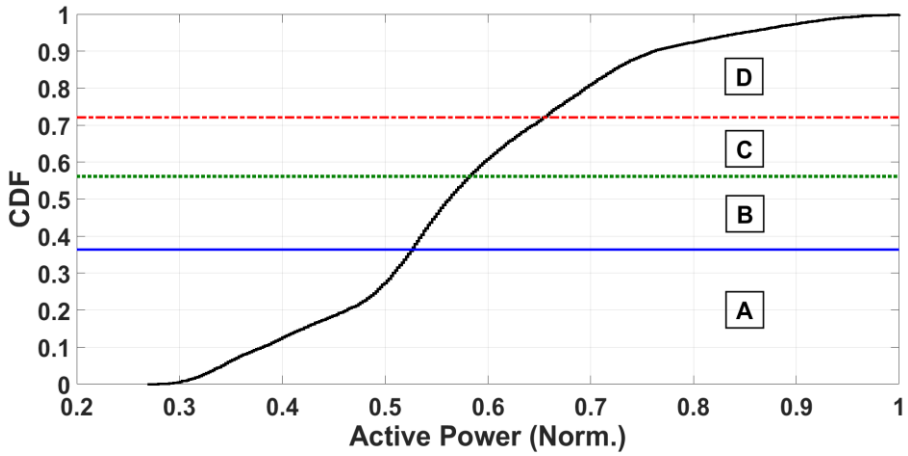


Figure 6.12: CDF of normalised active power and corresponding Areas-A to D

This approach can be modified to account for shorter periods of time, i.e. within months or seasons of the year, in order to more accurately define the corresponding areas for short and medium-term expectations from base, intermediate and peak requirements. This will also limit the range of peak demands (shown as Area-D in Figure 6.12), to more appropriate levels for the corresponding periods, as it is shown to be an overestimation when considering the whole year and accounts for ~30 % of the total load, i.e. short-term generation planning will not benefit from peak load expectations for the whole year. The results in Figure 6.12 are, nevertheless, valid representations of load distributions according to the distinctions made, for one-year's measurements. Results for reactive power are not presented due to the low-significance of the seasonal component and weak correlations with temperature levels, as well as low regularity in reactive power demands patterns, as discussed in Chapters 3 and 4.

Distinctions between base, intermediate and peak reactive power demands are, nevertheless, possible, but due to these irregularities, these loads are not as well defined as in the case of active power. Other considerations include the fact that variability is not necessarily seasonal, i.e. the difference between base loads and peak loads is not always distributed according to temperature-related and yearly-related cycles. This has also been discussed in Chapter 5, for purposes of GSP classification and customer-class disaggregation, where the variable portion of total demand was weighted against a "seasonality-metric" (e.g. R^2 of the *P-T* seasonal correlations, presented as Metrics 11&12 in Table 5.1).

6.3 Multiple-Regression Disaggregation

An obvious limitation of the "naïve" disaggregation approach presented in Section 6.2, is the fact that the resulting portions of the total demand are expressed as single or aggregate components, e.g. Areas C and D as shown in Figure 6.11, and not as disaggregated loads of particular load categories. Procedures aimed at a more detailed disaggregation are presented in the current section, based on multiple-regression analysis (also discussed in Chapter 4, Section 4.7), for which the generic model (for linear relationships) is in the form of:

$$y = \beta_0 + (\sum_{i=1}^{n} x_i \beta_i) + \varepsilon \approx f(X, \beta) \qquad (6.6)$$

where y is the measured active power demand, $f(X, \beta)$ is the estimated active power from the best fitted surface, X is a vector of independent variables x_i, for $i = 1,2..n$, where n is the selected number of independent variables, β_i are the corresponding beta-coefficients, β_0 is the y-intercept and ε are the error terms, or residuals. The analysis is performed on a per half-hour of the day basis, where the data-points are the weekdays of the year, excluding a 10-day period during Christmas which includes significant demand deviations due to the holiday season, as discussed in Chapter 4, Section 4.7. When the best fit functions are computed the beta-coefficients are retrieved and then, the estimated loads are in the form of:

$$y_i^E = \beta_i \cdot (x_i - x_{iB}) \qquad (6.7)$$

where y_i^E is the estimated load contribution attributed to explanatory variable - x_i, β_i is the corresponding beta-coefficient or gradient of independent variable x_i and x_{iB} is the base value of variable x_i. The selected independent variable(s) bases, as presented in Section 6.1, are discussed in the following subsections due to their dependence on the particular load categories which are targeted by the corresponding disaggregation models.

This approach is based on the assumption that within the context of multiple regression analysis, the individual beta-coefficients - β_i can be considered as partial, i.e. the

increase/decrease in active power corresponding to gradient β_i can be attributed to explanatory variable x_i, because the effects of the remaining explanatory variables are held constant [219]. Furthermore, and for each model, the resulting coefficients of determination (R^2) are presented in order to demonstrate that there is a satisfactory goodness-of-fit between the dependent and independent variables. In the instances when this assumption fails, there are considerably high levels of residuals and therefore the disaggregated loads can be considered as less accurate estimations of the specific load categories, i.e. (6.7) does not account for the errors and therefore the best-fit surfaces must have good modelling performances. All model inputs are in normalised values with respect to the maximum value in each dataset (3.3), so that the resulting gradients and base values are comparable and within limits of [0 1]. Three different models are considered:

Multiple Regression Model-1: Active power with reactive power and temperature, denoted as ***P-QT***. This model is used for the disaggregation of total demand into thermal heating loads and based on the fact that electrical heating elements can be considered as primarily/purely resistive in terms of their electrical characteristics and with unity power factor [220], [221].

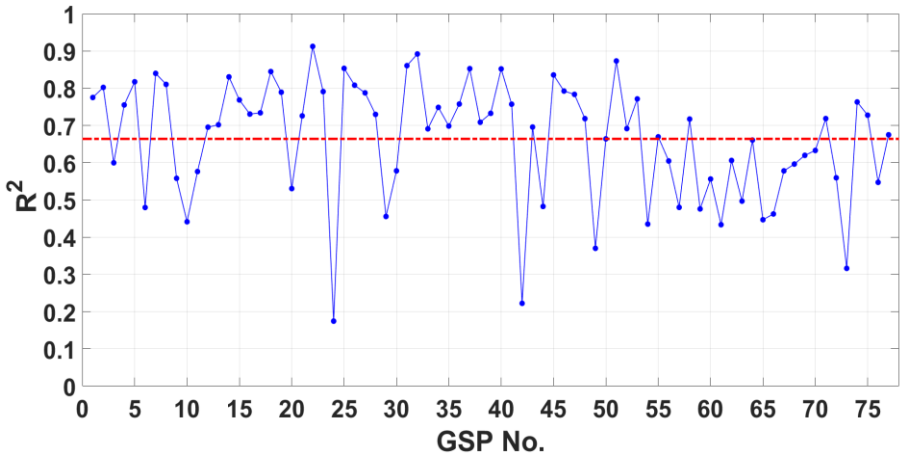


Figure 6.13: Model performances in coefficients of determination - R^2 for 77 GSPs, for *P-QT*

The model is aimed to account for the seasonal changes in active power demands that can be attributed to the corresponding changes in temperature, at constant reactive power levels. The overall model performances, as applied to measurements from 77 GSPs, are presented in Figure 6.13, in terms of the coefficients of determination - R^2.

Multiple Regression Model-2: Active power with reactive power, temperature and solar elevation angles, denoted as ***P-QTE***. The model is used for the disaggregation of total demand into thermal heating loads, as Model-1, but in this case the variations of solar elevation angles are also taken into account, thus allowing for a portion of the variability to be attributed to the

effects of illumination levels. The model performances for 77 GSPs are shown in Figure 6.14. The *P-QTE* model has better overall performance than the *P-QT* model, with an average R^2 of ~0.75 compared with an average value of ~0.65 for Model-1, as expected due to the inclusion of the additional explanatory variable (i.e. solar elevation angle). In both cases the goodness-of-fit varies among GSPs form ~0.2 to more than ~0.9. There is also an evident tendency of relatively higher R^2 values for the Scottish datasets (first 53 GSPs), compared with the English datasets (GSPs 54-77).

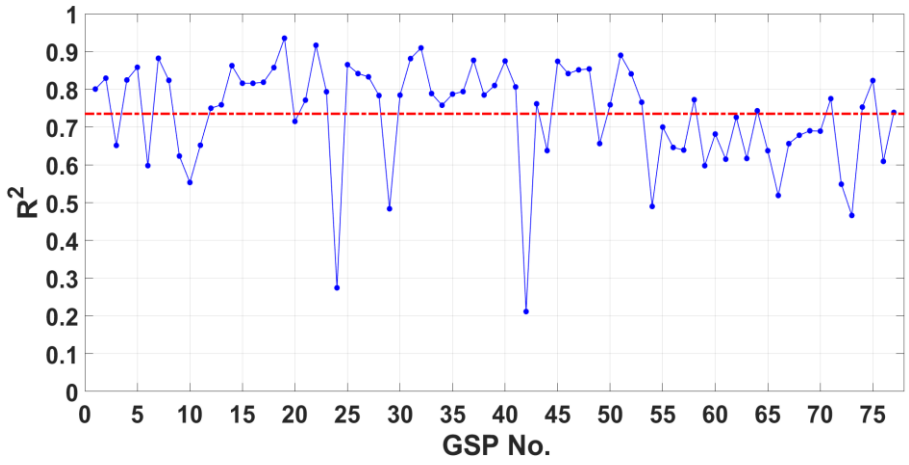


Figure 6.14: Model performances in coefficients of determination - R^2 for 77 GSPs, for *P-QTE*

Multiple Regression Model-3: Active power with temperature and solar elevation angles, denoted as ***P-TE***. This model is used for the disaggregation of total demand into lighting loads, based on the assumption that the portion of seasonal variations that is not associated with temperature can, at least to some extent, be attributed to the variability of lighting loads. Note that reactive power cannot be used in this case to enhance model performance, due to the fact that lighting appliances have varied electrical characteristics and therefore it cannot be assumed that the aggregated lighting loads operate at a single characteristic power factor level[21] [222]. The model is also used for the disaggregation of total demand into thermal cooling loads, for which reactive power also cannot be assumed to be kept constant during the operation of air conditioning systems [221]. The *P-TE* model performances are shown in Figure 6.15, with an average R^2 of ~0.65.

[21] Although it can be assumed that incandescent lamps have been used prior to 2009-2010, at least in the residential sector, the available demand datasets come from various sector-mixtures and from the period between 2007 to 2014.

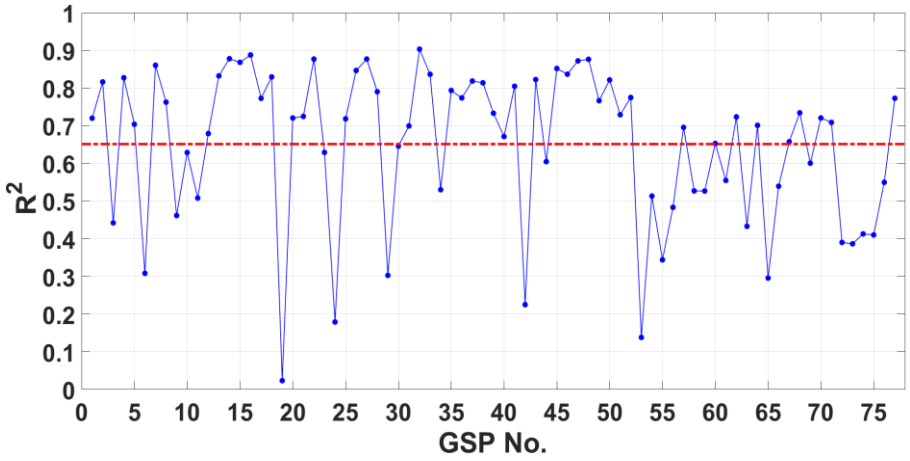


Figure 6.15: Model performance in coefficients of determination - R^2 for 77 GSPs, for *P-TE*

For each model and each disaggregated load-type, the estimated contributions are given as percentages of total measured demand, i.e.:

$$y_i^E(d,t)\ \% = \frac{y_i^E(d,t)}{y(d,t)} \times 100 \qquad (6.8)$$

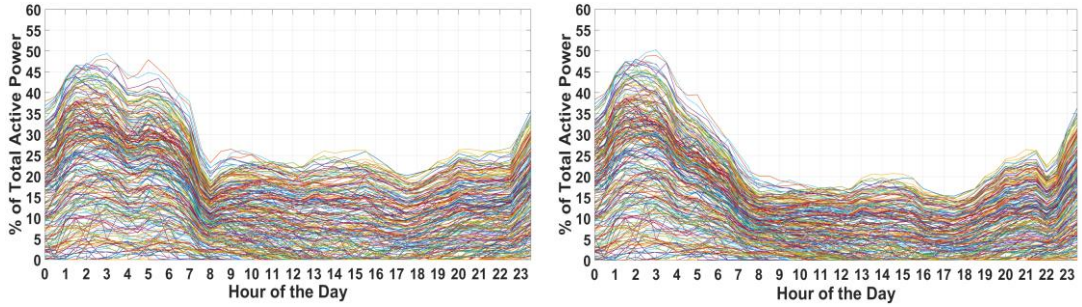
where (d,t) is the particular data-point at weekday – *d* and half-hour – *t*. All disaggregated loads are based on the analysis of 77 GSPs (for which data for *P*, *Q*, *T* and *E* were available) and the final percentages-contributions are represented by the 50th (median), 95th and 5th percentile values. Examples are also provided for particular GSPs, in terms of the seasonal and diurnal profiles of the disaggregated loads, to highlight the differences among the different models, as well as the differences between buses with different customer-class mixtures.

6.3.1 Thermal Heating Loads

For the disaggregation of total demand into thermal heating loads, models *P-QT* and *P-QTE* are used, following the previous discussion. Regarding the temperature base, i.e. x_{iB} as described in (6.7), the value of 15 C^o is selected. The choice is based on a comparison between temperature base values mentioned in literature (~15.5 C^o) and the results presented in Section 6.1.3, which showed lower thresholds, as least for certain periods of the day. Alternatively, and based on the same analysis, the threshold temperatures can be allowed to vary throughout the diurnal cycle according to Figure 6.6, however analysis has shown that the variable bases produce estimations that are, on average over all GSPs, approximately equal to the constant base estimations.

Figure 6.16 shows an example of the resulting thermal heating components as percentages of measured demand (6.8), for GSP-14, for all weekdays of the year in the diurnal perspective and for models *P-QT* in Figure 6.16 (a) and *P-QTE* in Figure 6.16 (b). The results indicate a

high presence of economy-7 type customers with thermal heating demands during night hours, reaching ~40 % to 45 % of total measured load, for the winter months. This is in agreement with results presented in Chapter 5, which showed high E7 percentages for GSP-14 at ~20 % of total consumption.

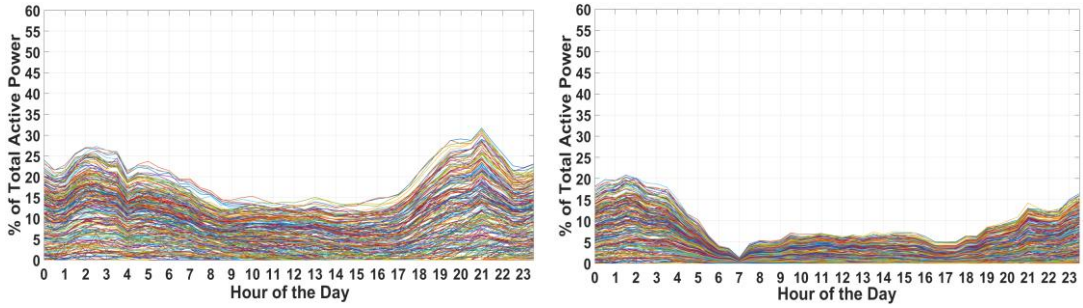


a) **b)**

Figure 6.16: An example of the disaggregated thermal heating loads for GSP-14: a) *P-QT* and b) *P-QTE*

The inclusion of solar elevation angles in the model brings the overall percentages down, throughout the day, apart for the night period between 23:30 to 04:00 hours, as shown in Figure 6.16 (b). This shows that a portion of the variability is now accounted by loads relating to solar irradiance levels (i.e. lighting loads) and particularly for the periods between 04:00 to 08:00 hours as well as during the mid-day period.

The flexibility of the disaggregation approach is further demonstrated in Figure 6.17 using the results from the *P-QTE* model, for GSPs-33 and 3.



a) **b)**

Figure 6.17: An example of the disaggregated thermal heating loads using model *P-QTE* for a) GSP-33 and b) GSP-3

GSP-33 (~62 % TR), can be assumed to have a higher percentage of direct electric heating, rather than economy-7 storage heating, indicated by the distinctive peak during the afternoon/evening period, i.e. between 17:00 to 22:30 hours. This could also be a result of increased demands from economy-10 customers, as E10 has reduced tariffs for the evening period, as well as for the night period. GSP-3, corresponds to mixture/commercial consumption (~50 % TR) and has lower overall percentages of thermal heating loads, with

peaks during the night period that can be attributed to the lower percentages of residential economy-7 customers (as a percentage to total measured demand – the actual consumption is not lower than that shown for GSP-33). The results are in agreement with the results presented in Chapter 5, which show relatively low E7 contributions of ~10-15 % for GSPs-3 and 33.

Figure 6.18 presents the synoptic results for the thermal heating contributions, in the seasonal perspective and considering the results from all 77 GSPs, based on the *P-QT* and *P-QTE* models. The maximum range, for the winter period, is between ~15 % to ~25 %, for both methods, while minimum contributions are concentrated in the period between mid-June and mid-September at levels between approximately 0 % to 5 %.

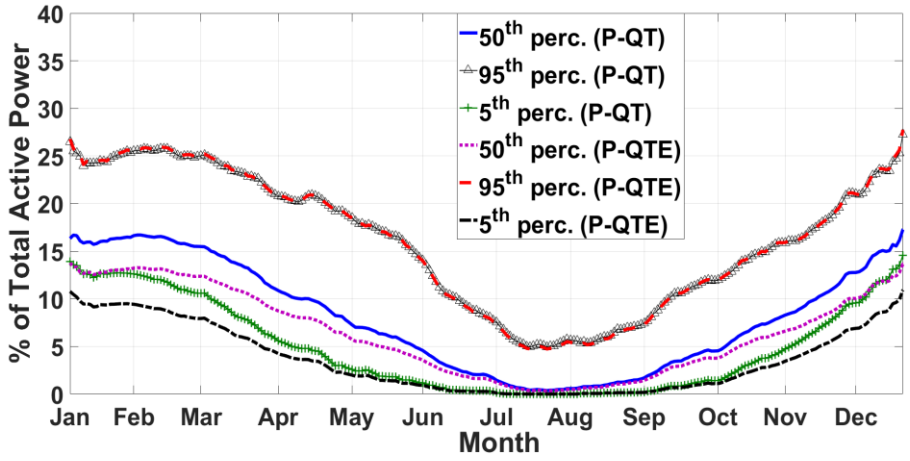


Figure 6.18: Estimated contributions from thermal heating loads based on the analysis of 77 GSPs, for models: *P-QT* and *P-QTE*

Differences between the two models are evident for the median and 5th percentile values, with lower estimations for the *P-QTE* approach, indicating as mention above, that a portion of the overall variability is now attributed to loads associated with solar irradiance levels. Note that for individual GSPs the percentages can be outside the limits shown in Figure 6.18, as it is the case for primarily residential GSPs, an example of which was presented in Figure 6.16.

Figure 6.19 presents the consistency of the two models among the 77 GSPs used in the analysis. Figure 6.19 (a) shows the empirical CDF for the mean difference between *P-TQ* and *P-TQE*, as a percentage to total active power demand. This is shown to be below 6 % for all GSPs and below 3 % for 90 % of the GSPs. Figure 6.19 (b) shows the correlation (in terms of R^2 values), for the mean diurnal thermal heating contributions for each GSP between the two models, i.e. correlations of resulting daily profiles, which is at levels above 0.9 for approximately 60 % of the sample and above 0.8 for approximately 80 % of the sample. The diurnal distribution of the differences between the two models is shown in Figure 6.19 (c), as percentages of total active power. These are the averaged per half-hour differences from all

analysed GSPs. Note that the difference is positive as the results from model *P-TQE* are subtracted from the results of model *P-TQ*. The differences are concentrated during the morning period, i.e. between 04:00 to 10:00 hours, with a smaller peak during the period between 16:00 to 19:00 hours and are below 1 % of total demand for night hours.

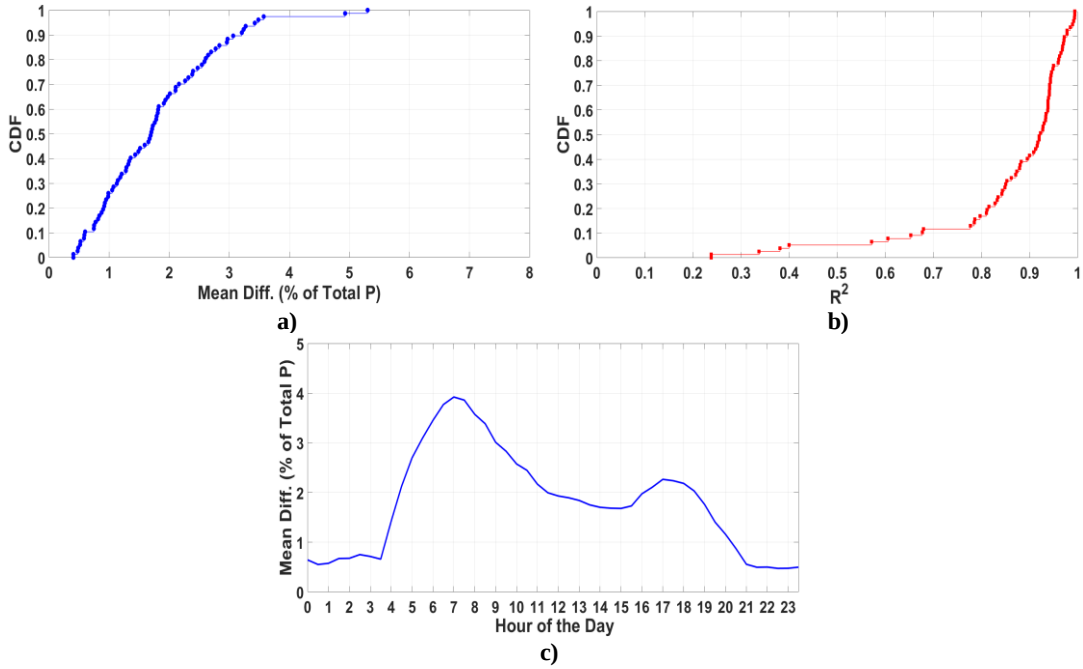


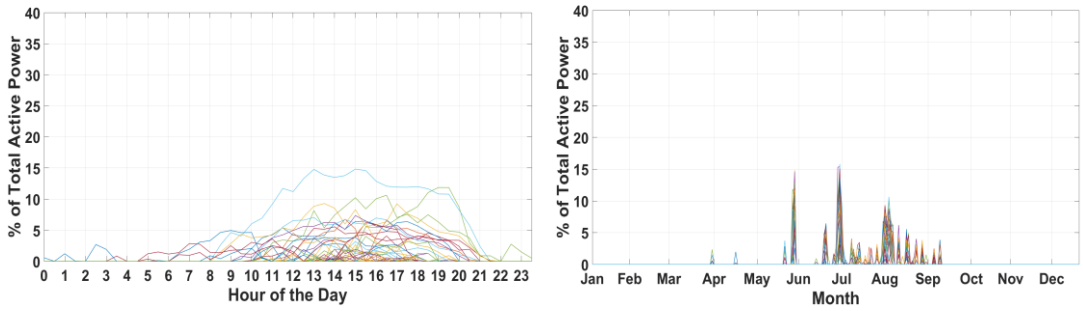
Figure 6.19: Consistency of estimations between models *P-TQ* and *P-TQE*: a) CDF of mean % difference, b) R^2 of diurnal profiles and c) diurnal distribution of the differences

6.3.2 Thermal Cooling Loads

Thermal cooling load estimations are computed based on model *P-TE*, for the reasons discussed in the introduction of Section 6.3. No significant cooling loads can be reported for cooling base temperatures of 22, 20, and 18 C°, while any further adjustment of the corresponding base below these temperatures can be considered unrealistic, i.e. demands for thermal cooling loads (fans and air-conditioners) are not expected at lower temperatures. This assumption is supported by the cooling base reported for the UK, which is at levels above 20 C°. The results are in agreement with the moving-window regression analysis presented in Chapter 4, Section 4.8, which showed that although, for some GSPs, positive beta-coefficients for the active power-temperature relationship were estimated during the summer period (indicating increase in active power demands with an increase in temperature levels), these had low statistical-significance, as quantified by the low coefficients of determination for the corresponding periods.

There are, however, some minimum cooling load contributions for a number of GSPs and for restricted diurnal and seasonal periods, an example of which is presented in Figure 6.20, for

GSP-11. These are concentrated primarily during the summer months (as expected) and during the diurnal period between 10:00 to 20:00 hours, peaking during mid-day and early afternoon hours, indicating that they are most likely related to the commercial sector (e.g. retail shops, offices, etc.). For the example GSP presented in Figure 6.20, contributions from the industrial and commercial sectors have been estimated at ~ 40 % of the total consumption (as presented in Chapter 5).



a) b)

Figure 6.20: An example of the disaggregated thermal cooling loads using *P-TE* model for GSP-11: a) diurnal and b) seasonal perspectives

The synoptic results for all analysed GSPs are presented in Figure 6.21, showing maximum thermal cooling loads below 2 % of the total measured load for the period. With respect to the total annual demands, analysis has shown that the contributions from thermal cooling loads does not exceed 1 %, for all 77 GSPs.

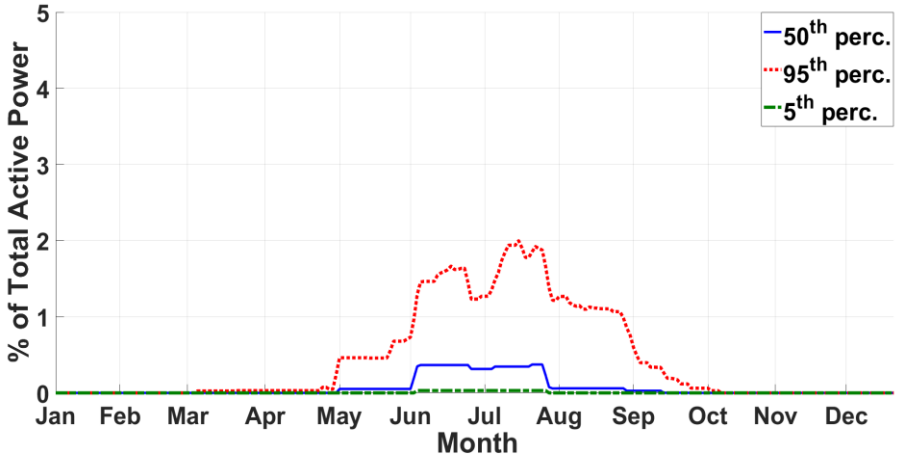


Figure 6.21: Estimated contributions from thermal cooling loads based on the analysis of 77 GSPs, using the *P-TE* model

6.3.3 Lighting Loads

As discussed in Section 6.1.4 and unlike thermal (heating and cooling) loads, the active power-solar irradiance relationship is weak and non-linear and does not have clear identifiable threshold values, or "turning-points", which can be used to determine contributions from

lighting loads. Unlike thermal demand, demand for electrical lighting is always present throughout the year and for all hours of the day (e.g. indoor spaces with no natural-lighting) and particularly when considering aggregated demands from MV-GSPs, in which case it is impossible to determine periods of zero (or near-zero) lighting loads in the system. The variety of lighting appliances with different electrical characteristics means that reactive power variations cannot be taken into account in the same fashion as for thermal heating loads, i.e. in order to determine changes due to mostly/purely resistive loads (demanding only *P*, not *Q*). Another important limitation is the absence of detailed validation data in the seasonal and diurnal perspectives that can be compared against the results presented in this section, so that the procedures can be calibrated against targeted estimations. Available percentages from literature are concentrated on household consumption and on total national and sub-national consumptions but, as shown in Chapter 5, the majority of MV-substations include demands from various customer-classes. In the same context, there are also no detailed load profiles regarding public street lighting.

The presented methodologies are based on the *P-TE* multiple regression model as presented in the introduction of Section 6.3. However, and due to the complications discussed above, the disaggregation approach is further divided into two sub-models, each of which includes further modifications to the original *P-TE* procedure and are denoted as Models A and B. The results, as in the case of thermal loads, are presented in the seasonal and diurnal perspectives for example GSPs and as final estimations of the seasonal percentages, based on the analysis of 77 GSPs, using percentile values.

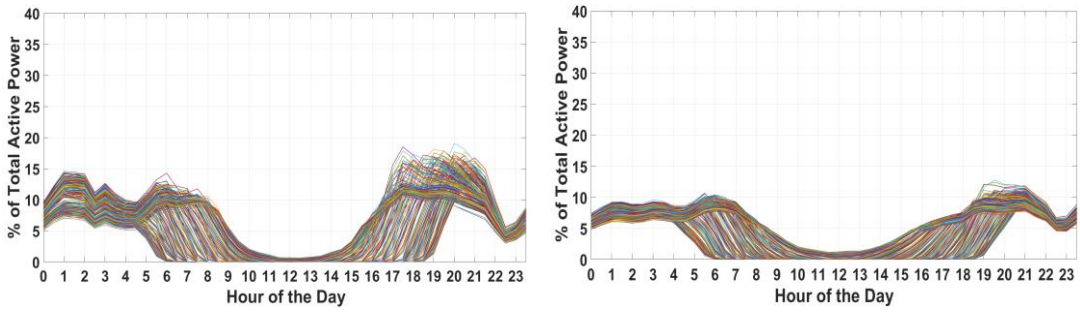
Model-A: a constant solar threshold value is used, at a normalised level of 0.27, denoted by the black-dotted line in Section 6.1.4, Figure 6.9. The topocentric solar elevation angle is used as the second explanatory variable, the first being temperature. This is allowed to vary during night hours (as discussed in Chapter 4, Section 4.2) so that the regression models are not compromised by the presence of zero values during early morning and evening hours, when solar elevation angles are non-zero only for short periods of the year, a fact that results in poor regression models at the corresponding half-hours. However, during the load estimation phase, i.e. (6.7), a constraint is set so that:

$$x_i(d,t) = 0 \ \ for \ \ x_i(d,t) < 0 \qquad (6.9)$$

and therefore lighting load estimations are not allowed to vary for half-hours - *t* at weekdays - *d* that have negative solar elevation angles, i.e. the regression models are evaluated on the topocentric elevation angles (allowed to vary during night hours), but the lighting load estimations are constraint during those periods, according to (6.9). Removing this constraint

would allow a large portion of load variability during night hours to be attributed to lighting demand, which would then produce unrealistic results for the corresponding periods, i.e. lighting loads are relatively constant during the night, throughout the year, but the collinearity between temperature and solar elevation would result in higher lighting load contributions during the corresponding periods. Another consideration regards the fact that the estimated loads can be negative when $(x_i - x_{iB})$ is positive, i.e. when solar elevation angles are above the selected threshold value. In this case the negative estimations are set to zero, as it is assumed that no lighting loads are required for solar elevation levels above the threshold.

An example of the results is shown in Figure 6.22, for GSP-14 and GSP-3. The presence of zero values is due to the fact that the procedure can only account for the seasonally variable loads and not for the seasonally constant loads. Furthermore, and as previously mentioned, the approach is based on the assumption/simplification that above the threshold values no lighting loads are required.



a) **b)**

Figure 6.22: An example of disaggregated lighting loads using Model-A for: a) GSP-14 and b) GSP-3

The shifts in the peaks during the evening hours (and particularly for the residential GSP-14) are in agreement with the result for the rate of change of active/reactive power, as illustrated in Chapter 3, Section 3.8, which showed that during the late evening, demands increase later during the summer months (highest excursion during June) and earlier during the winter months (in December). This seasonality was also demonstrated in Chapter 3, Section 3.7, by the periods of occurrence of peak demands in the system, which also showed seasonal variability. There are however, for both GSPs, unrealistic load patterns, including the decreasing trends for several weekdays of the year between 04:00 to 10:00 hours, a period at which increasing demand for lighting is expected throughout the year, as well as the night peak shown for GSP-14, between 01:00 to 02:00 hours. There are also distinctively low contributions (close to zero) during the mid-day period, i.e. working/office/retail hours, which may indicate low seasonal variability of lighting loads during this time.

The final seasonal estimations, based on the analysis of 77 GSPs are presented in Figure 6.23, in median, 95th and 5th percentile values. The results vary between ~ 5 % to ~14 % for December and January and between 0 % and ~ 10 % for mid to late June, coinciding with the summer solstice, as expected.

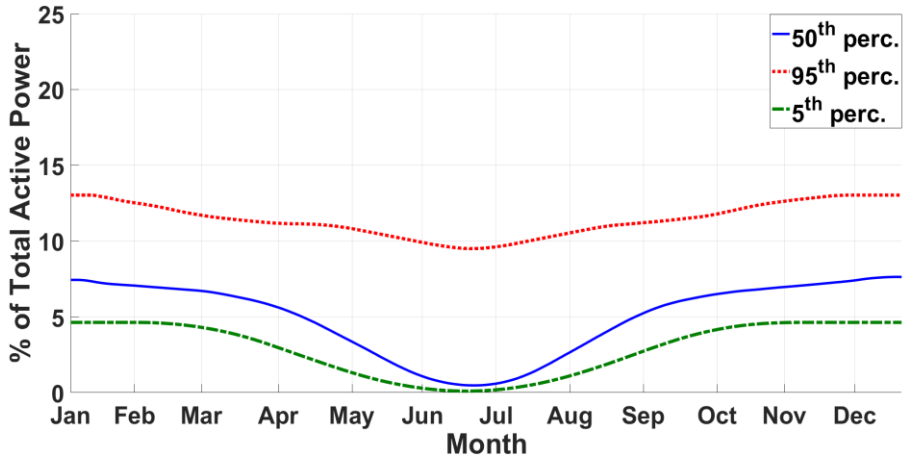
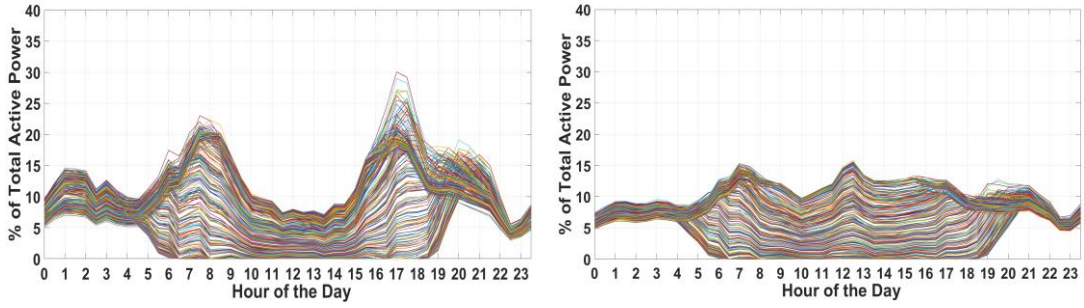


Figure 6.23: Estimated contributions from lighting loads (seasonally variable) based on *P-TE*, Model-A

Model-B: is based on the same approach as Model-A, modified by using a variable solar elevation base that is allowed to change with respect to the half-hour of the day, as presented in Section 6.1.4, Figure 6.9 (denoted by the blue-solid line), but which is not allow to vary below the mean threshold level of 0.27, in order to account for lighting loads during the night hours. Examples are shown in Figure 6.24, for GSPs-14 and 3, in the diurnal perspective.



a) **b)**

Figure 6.24: An example of disaggregated lighting loads using Model-B for: a) GSP-14 and b) GSP-3

The results for Model-B better match the assumed diurnal patterns of lighting load consumption, also found in literature (e.g. from household consumption statistics in [106]), with distinctive peaks during morning and afternoon to evening hours for the residential GSP, in Figure 6.24 (a), and higher contributions during mid-day for the commercial GSP, in Figure 6.24 (b). The results therefore indicate that using a variable solar base produces estimations that better resemble the inherent characteristic profiles of the corresponding GSPs.

However, as with Model-A, the approach cannot account for portions of lighting load that are below the seasonal variations limit (as described in Section 6.2). While the range of percentages appears excessive for lighting loads, when considering the seasonal component only (i.e. percentages in literature vary between 5-20 % for the total contributions in the residential sector [106]), it should be noted that the presented percentages are with respect to measured demands at the corresponding weekdays and half-hours, i.e. percentages tend to increase for periods of lower overall demand. When the synoptic results are considered, based on the percentile contributions from all available GSPs, the estimations fall within more reasonable limits. These values are presented in Figure 6.25.

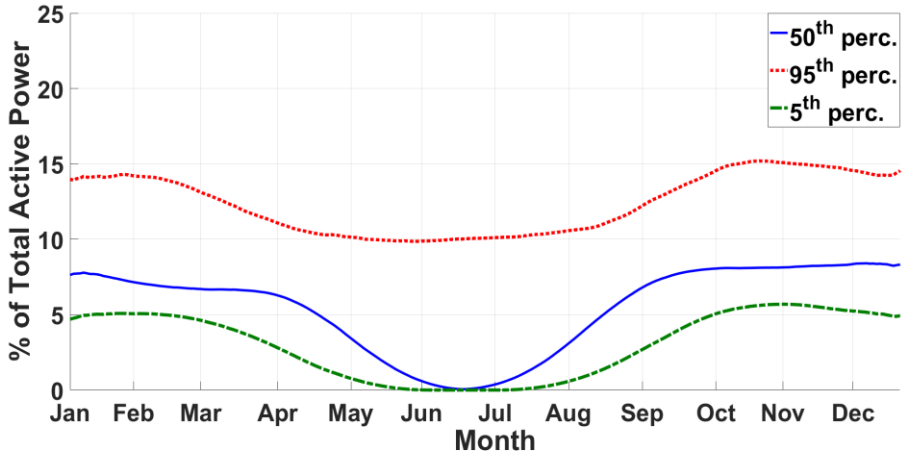


Figure 6.25: Estimated contributions from lighting loads (seasonally variable) based on *P-TE* Model-B

Both models performed well in capturing the (assumed) seasonal variability of lighting loads, with the minimum and maximum contributions estimated during the summer and winter solstices, respectively (shown in Figures 6.23 and 6.25). These correspond to the periods of maximum and minimum available ambient sunlight. Furthermore, minimum lighting loads do not coincide with minimum thermal heating loads, as presented in Figure 6.18, with the latter found approximately one month later, at the end of July - beginning of August. This observation is in agreement with the seasonal patterns of temperature and elevation angles, as presented in Chapter 4, Figure 4.33, which showed that the two parameters are out of phase by approximately 1.5 months, and thus a similar difference in the minimum disaggregated heating/lighting loads was expected.

As Model-B produces realistic/expected consumption patterns, i.e. morning and evening peaks for the residential GSP and midday peaks for the commercial/mixture GSP, a base lighting demand can be added to the results in order to produce estimations of the total lighting load. The added base needs to be constant with respect to the seasons (as the seasonal variability is

captured by Model-B), but variable according to the hours of the day, as demand for lighting is sensitive to the different diurnal periods, as well as with respect to customer-sectors (i.e. consumption profiles). For example, for the residential GSP, a 5 % base can be added and similarly a 15 % base can be added for the commercial/mixture GSP (as higher, seasonally-constant lighting demands are expected from the non-domestic sector), [71], [106] and [111].

Examples, for GSPs 14 and 3 are presented in Figure 6.26. The results are calculated by adding the mean contributions to total demand from Model-B (Figure 6.24) to the, assumed, base lighting load demands. The bases correspond to the selected percentages, i.e. 5 % for GSP-14 and 15 % for GSP-3, multiplied by the mean demand per half-hour of the day.

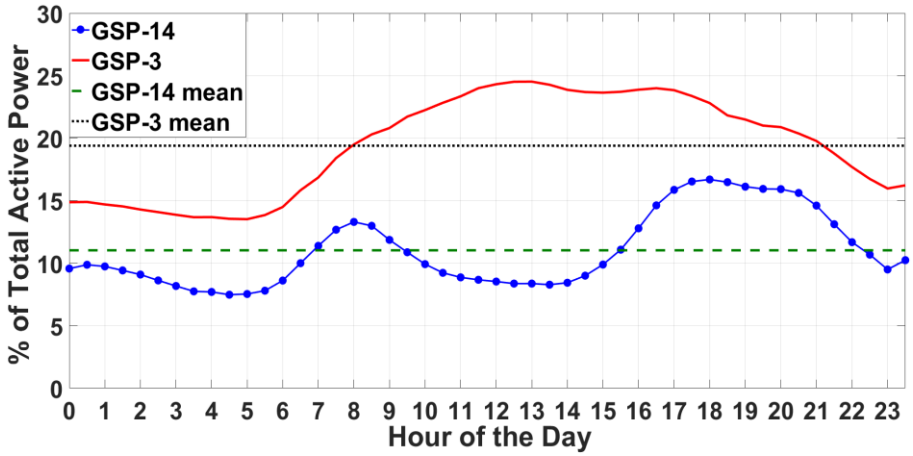


Figure 6.26: Mean lighting load contributions to total active power demand, for GSPs 14 and 3

These final adjustments are based on some reasonable but nevertheless precarious assumptions. In particular, the percentages of the added base-contributions are chosen to match assumed total demands for lighting, which might not be accurate and are not tailored for the specific customer-classes supplied by the corresponding GSPs. Furthermore, these percentages are added according to the mean demands per half-hour, e.g. for GSP-14, a 5 % increase during the evening corresponds to more load than a 5 % increase during mid-day, and the adjustment is therefore based on the assumption that lighting load demands are linearly correlated with total demand per half-hour. The estimations can be improved by informing the final base adjustments with known disaggregated percentages from SM-data, corresponding to specific customer-classes and to the same locations as the MV-data.

6.4 Data Transformations and Power Factor Analysis

This section introduces a novel methodology developed for the disaggregation of active power into components of seasonally-variable thermal and non-thermal loads. The approach is based on the analysis of the active power - apparent power relationship, or power factor (PF), which can be expressed as:

$$PF(d,t) = \frac{P(d,t)}{S(d,t)} = \frac{P(d,t)}{\sqrt{P(d,t)^2 + Q(d,t)^2}} \quad (6.10)$$

where *P* and *Q* are the measured active and reactive power demands and *S* is the corresponding apparent power, at half-hour - *t* and weekday – *d*, excluding a 10-day period during the Christmas holidays, as in the previous section. By rearranging (6.10) it is possible to construct (counterfactual) active power values, at constant power factor levels - *pf*, that correspond to the actual/measured reactive power, in the form of:

$$P_E(d,t,pf) = \frac{pf}{\sqrt{1-pf^2}} \times Q(d,t) \quad (6.11)$$

where *pf* is then allowed to vary between [0 1] in iterative steps of 0.01 (or otherwise), so that for each half-hour of the day arrays of "expected" active power values - P_E are calculated at each *pf* level. Since these values are mapped from a linear function of reactive power and a gradient (i.e. $\frac{pf}{\sqrt{1-pf^2}}$), they are perfectly correlated with $Q(t)$ and thus the remainder between actual measured active power and P_E is assumed to be related to electrical resistive loads (i.e. the variations of $Q(t)$ are assigned to the corresponding $P_E(t,pf)$). The difference is given by:

$$P_D(t,pf) = P(t) - P_E(t,pf) \quad (6.12)$$

where $P_D(t,pf), P(t)$ and $P_E(t,pf)$ are arrays with weekdays of the year as data-points. Single axis transformations are essentially considered in the form of x-axis rotations, so that the remainder P_D gradually decreases for increasing power factor levels - *pf*. However, due to the inherent variability of the actual active power demand and the gradients of the transformation axes, the gradual decrease in active power values is not uniform among the weekdays of the year, at constant half-hours. As a result, the corresponding seasonality of the P_D datasets changes for different *pf* levels and this is exploited in the second step of this analysis, where these values are correlated with measured temperature and reactive power.

While the analysis is presented in terms of power factor levels, from a mathematical perspective the described transformations may not be necessarily restricted to *pf* changes and can be expressed as functions of simple gradient ascents, such that:

$$P_E(t,b) = b \times Q(t) \qquad (6.13)$$

for variable values of b within some chosen range, as in the case of *pf*. In fact, due to the denominator term in (6.11), i.e. $\sqrt{1-pf^2}$, there is an increasing "gap" between the transformed axes for increasing *pf* values, as it can be seen in Figure 6.27. This means that, for constant iteration steps, the resolution of the analysis decreases as the reference power factors approach unity. For very small iteration steps (e.g. Δb or Δpf of 0.005 or smaller) and for relatively small *pf* (or b) levels, the two methods produce similar results, however, the two quickly diverge for higher values. Because actual demand levels are usually restricted within power factors of ~0.95 to 1 and because the P_D values of interest correspond to demand differences which are measured below the lower limits (i.e. P_E lines below measured demand), the use of power-factor transformations is considered acceptable. Note that there is also an infinite number of possible functions that can determine P_E (6.11) and thus P_D(6.12) which may be non-linear, i.e. these are not restricted to linear-polynomial functions, nor to polynomial functions for that matter. In fact, it is reasonable to assume that other transformation functions maybe proven more appropriate for the proposed analysis.

The procedure, as described so far, is illustrated in Figure 6.27, using *P*-*Q* measurements (non-normalised) from GSP-14 at 16:00 hours. An example of the estimated quantities is shown for a particular weekday of the year (intersections of red-dashed lines).

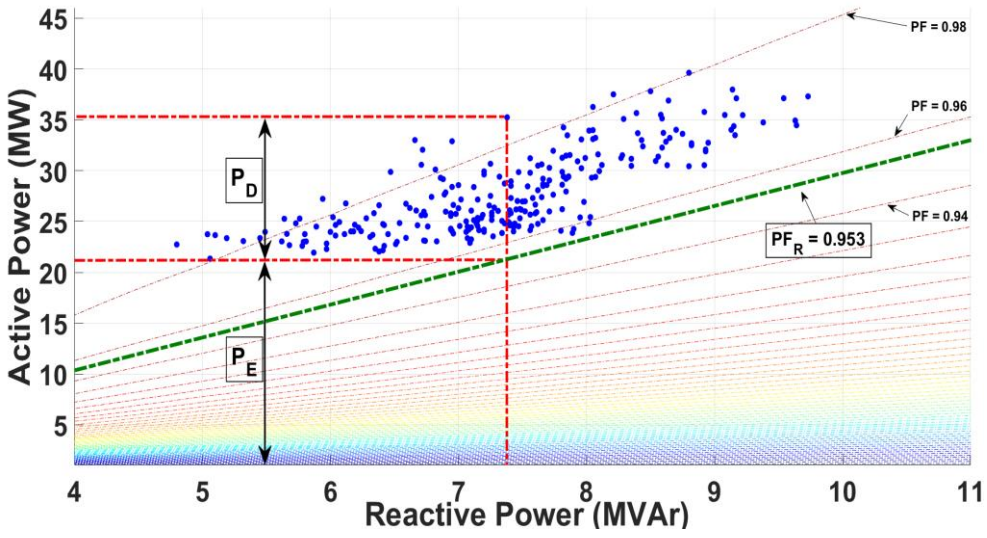


Figure 6.27: Illustration of the approach for determining the P_E and P_D portions of total active power, for GSP-14 at 16:00 hours

Marked on Figure 6.27 is also a specific power factor level of 0.953, denoted as reference power factor - PF_R. This corresponds to the power factor which produces a P_D dataset that minimizes the correlations with the measured reactive power Q or alternatively the reference

power factor can be selected at the point at which P_D maximizes the correlations with measured temperature - T. Ideally, the corresponding maximization/minimization transformations will coincide, so that the resulting P_D at reference power factor PF_R (or reference gradient - b_R) will simultaneously do both, thus correctly representing thermally-dependent resistive types of loads. Examples of the resulting correlations in squared correlation coefficients (4.3), for 48 half-hourly periods are shown in Figure 6.28 (a) for P_D with reactive power and in Figure 6.28 (b) for P_D with temperature, for GSP-31.

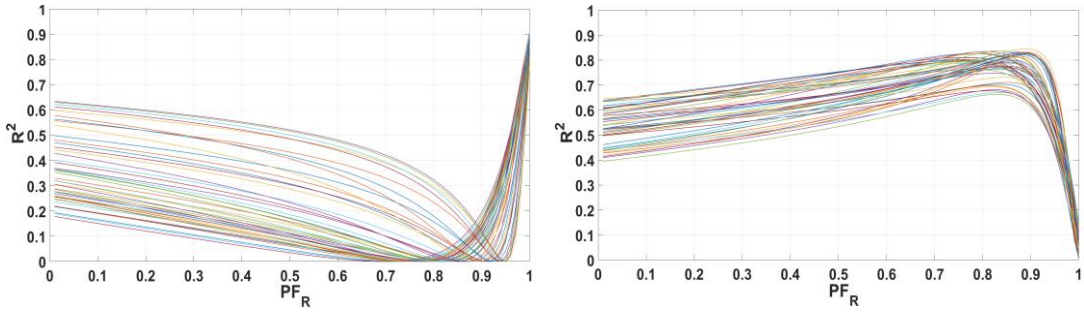


Figure 6.28: Examples of the resulting P_D correlations with: a) reactive power - Q and b) temperature - T, for 48 half-hours, GSP-31

There is a general tendency of decreasing correlations of P_D with reactive power for increasing reference power factors and similarly, a tendency of increasing correlations with temperature. For this particular GSP, there is a very good "agreement" between the points of minimization/maximization, as it is shown in Figure 6.29, for all 48 half-hours of the day.

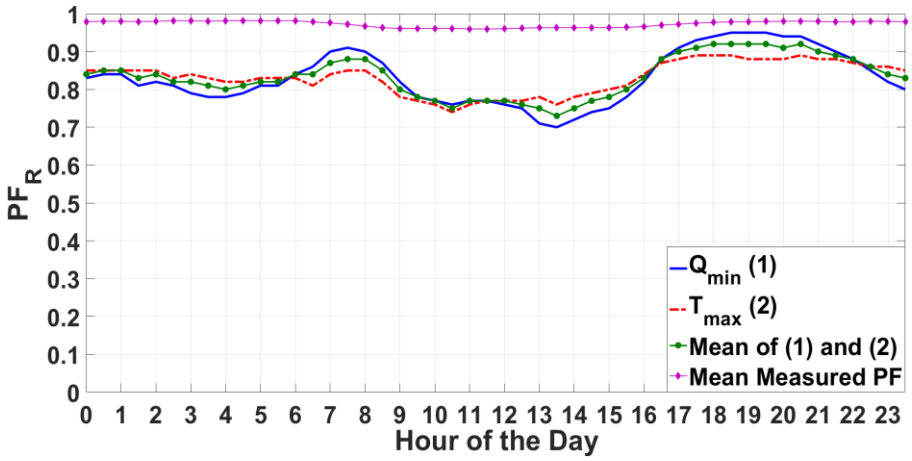


Figure 6.29: Resulting reference power factors - PF_R, for 48 half-hours, GSP-31

Marked on Figure 6.29 are also the mean PF_R of the minimization/maximization power factors, as well as the mean power factor values calculated directly from the measured P-Q values, in order to demonstrate that the transformations do not exceed the limits of actual active/reactive power levels, i.e. P_E is not larger than P and thus P_D is positive. These

conditions are not always met and the "ideal" scenario shown in Figure 6.29 is not representative of all GSPs, at all half hours of the day. In particular, for a number of GSPs, there is a low correlation and high deviation between the T_{max} and Q_{min} reference power factors, for the 48 diurnal periods. When the minimization/maximization reference power factors do coincide, the estimated loads P_D have improved correlations with temperature and decreased correlations with reactive power and thus the assumption that follows is that the portion of active power given by P_D is temperature related (to a higher degree than the original *P*) and is also mostly/purely resistive (without *Q* demand), which can be considered as the characteristic "signature" of thermal electrical heating loads.

The results for 77 GSPs are shown in Figure 6.30 using: a) the Pearson's correlation coefficient (4.3) that quantifies the covariance of the diurnal patterns of the two reference power factors and in b) the mean absolute error to quantify their differences, i.e. sum of absolute differences averaged over the 48 half-hours of the day, per GSP:

$$MAE = \frac{1}{48}\sum_{t=1}^{48}|PF_R(Q_{min}(t)) - PF_R(T_{max}(t))| \qquad (6.14)$$

Agreement between moderate/strong correlations and low MAE is shown for approximately 30 % of the total number of GSPs. This is based on (arbitrarily) chosen correlation levels of 0.6 (satisfied by ~40 % of the GSPs) and MAE levels below 0.3 (satisfied by ~60% of the GSPs).

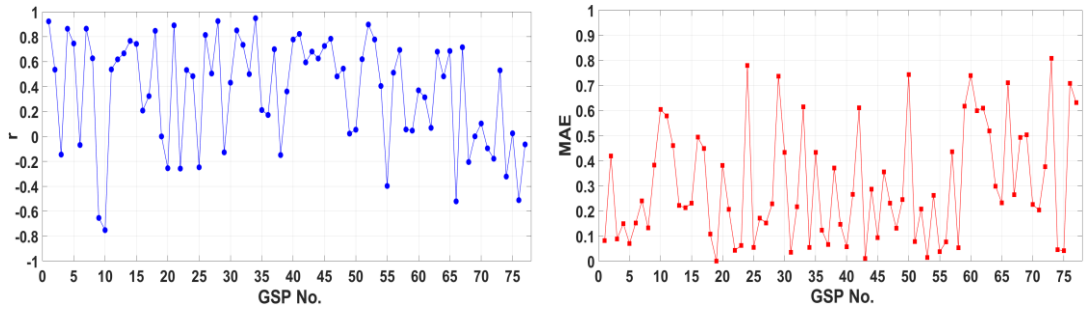


Figure 6.30: a) Correlation coefficients – *r* and b) mean absolute error – MAE, between the estimated Q_{min} and T_{max} reference power factors - PF_R, for 77 GSPs

In the instances where the PF_R for T_{max} and Q_{min} don not follow similar diurnal patterns and/or when their values largely deviate, the following justification can be given: at a specific *t*, if PF_R of Q_{min} is higher than PF_R of T_{max}, more load can be attributed to the total temperature related portion, however not all of it can be assumed to be mostly/purely resistive. This is particularly true during afternoon to evening hours for residential GSPs when occupancy related loads, lighting load, etc. are present. These loads are not mostly/purely resistive, but typically have seasonally variable components. This can be seen in Figure 6.29,

between 17:00 to 21:00 hours. Similarly, a primarily commercial GSP potentially includes a larger percentage of inductive/capacitive loads during the mid-day period (e.g. motors, power electronic, lighting loads, etc.) and thus the reference power factor for Q_{min} would be higher (during those half-hours) than T_{max}. Note that in all cases, and as it can be seen in Figure 6.27, the higher the PF_R the lower the percentage of the resulting P_D (with respect to total measured aggregate P) and thus less load is assigned to the corresponding load categories.

In the final step, the seasonal variations of active power demands are separated between the thermal (heating) components and the non-thermal components, denoted respectively as P_{TH} and P_{NTH} and calculated using:

$$P_{TH}(t) = \left(\frac{P(t)-P_B(t)}{P(t)}\right) \times P_D(t, PF_R) \quad (6.15)$$

$$P_{NTH}(t) = P(t) - (P_B(t) + P_{TH}(t)) \quad (6.16)$$

where $P(t)$ is the measured active power and $P_B(t)$ is the per half-hour seasonal base load, as calculated in Section 6.2 and illustrated in Figure 6.10 by the green-solid line. Equations (6.15) and (6.16) are based on simple assumptions relating to the presented axis transformations as well as on a heuristic approach after examination of the resulting P_D datasets. The P_D datasets are essentially overestimations for the diurnal periods when most of the seasonal variability can be attributed to thermal heating loads, because Q and T correlations can be minimized/maximized at low reference power factors. Similarly, for periods of low contributions of thermal-resistive loads the results are underestimations and often include negative values. In both cases the seasonal periodicity of the resulting P_D (per half-hour) is shown to be reasonable for heating demand and the problem is the overall level of these estimations. In the same context, simply removing the base - P_B produces underestimations and in many cases, non-zero results only for the diurnal periods of high heating demand. The proposed adjustment in (6.15) estimates thermal heating loads as the product of $P_D$values and the ratio of the seasonally variable load (Section 6.2) to the total load. It therefore restricts the values of P_{TH} within the limits defined by the actual recorded demand and the minimum recorded demand, at the specific half-hour of the day (over all weekdays). This adjusts both overestimations and underestimations and, assuming that the seasonality defined by P_D is correct, the results properly represent the desired load category. The remaining portion of the seasonal demand is then attributed to the non-thermal loads P_{NTH}, using (6.16).

An example of the correlations of the estimated loads with measured reactive power and temperature are shown in Figure 6.31, in order to demonstrate that these have improved correlations with temperature and decreased correlations with reactive power, in the case of

P_{TH}. Conversely, the analysis shows improved correlations for P_{NTH} with reactive power and decreased correlations with temperature. Since the analysis is based on transformations of *P-Q*, the changes in correlations are more pronounced for reactive power than for temperature. For reference, the initial *P-Q* and *P-T* correlations are also shown.

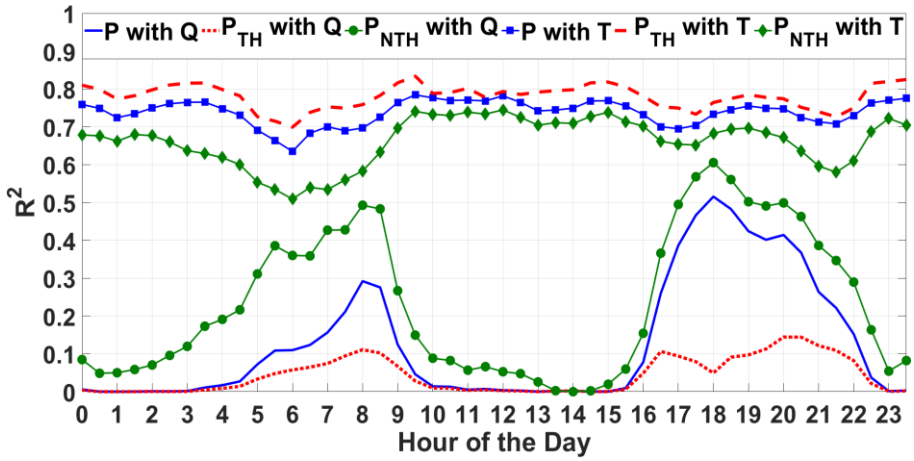


Figure 6.31: Correlations of measured and estimated loads (*P*, P_{TH} and P_{NTH}) with measured reactive power - *Q* and temperature - *T*

The choice of a reference power factor - PF_R in (6.15) is also not straight-forward and the considerations previously discussed regarding the correlations and deviations of T_{max} and Q_{min} should be taken into account, especially when the interest is concentrated on demands from individual GSPs and not on overall grid characteristics, as each individual GSPs generally have different levels and diurnal patterns for the resulting T_{max} and Q_{min} reference power factors.

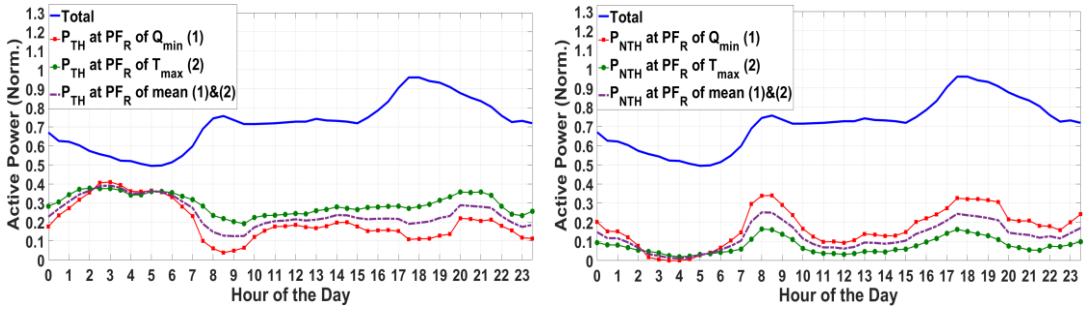


Figure 6.32: Examples of estimated loads for a winter day (day of peak demand), GSP-14, for: a) P_{TH} and b) P_{NTH} loads

An example of the differences in estimated loads is shown in Figure 6.32, for P_{TH} in (a) and P_{NTH} in (b), for a single winter day in normalised values (3.3), for GSP-14. In each plot, the estimations based on T_{max} and Q_{min} are presented, as well as an estimation based on the average of the two. The resulting loads from T_{max} and Q_{min} converge during the night between

02:00 to 06:00 hours, when the seasonal portion of the load can be mostly attributed to resistive thermal heating loads (for this specific GSP, also presented in the disaggregation approach of Section 6.3). During the same period, the corresponding P_{NTH} load drops close to zero and peaks during the morning between 07:00 to 10:00 hours and during the evening between 17:00 to 20:00 hours.

Figure 6.33 shows the results for the same GSP, in the seasonal perspective and for two characteristic hours of the day, i.e. at 17:00 hours in (a) and at 03:00 hours in (b). The loads presented in this figure are calculated with respect to the Q_{min} and therefore the disaggregation is based on minimizing the correlations with reactive power.

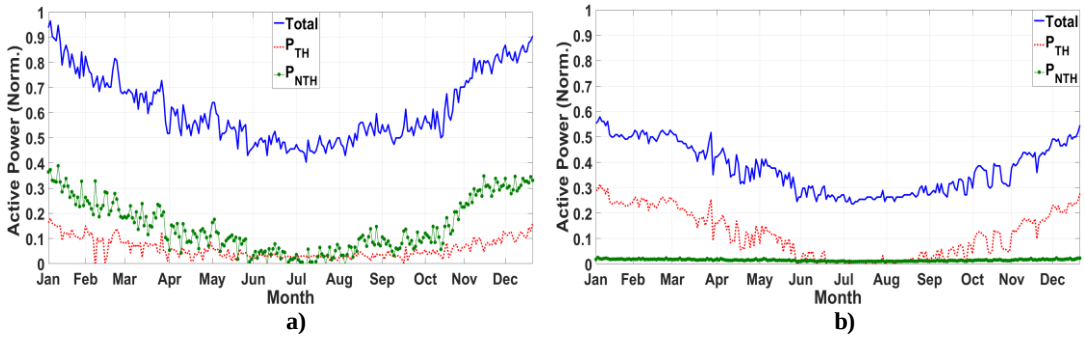


Figure 6.33: Examples of estimated loads P_{TH} and P_{NTH}, for GSP-14 at a) 17:00 hours and b) 03:00 hours

P_{NTH} loads are higher than P_{TH} loads during the afternoon, shown in Figure 6.33 (a), when there is a higher percentage of seasonally-variable demand that cannot be attributed to space heating, as discussed before. The reverse is true during night hours, in Figure 6.33 (b), when the P_{NTH} loads drop to zero and the demand for heating accounts for (almost) 100 % of the seasonal variability of active power. Since P_{NTH} loads are more strongly correlated with the variations of reactive power (Figure 6.31), inferences can be made about the load categories which they represent. These are assumed to include the seasonally variable portions of lighting loads, loads related to the use of electronic devices such as audio-visual equipment and ICT (PCs, monitors, networking, etc.) and in general non-unity power factor loads that can be affected by weather related changes (temperature and solar irradiance) and changes in occupancy levels.

Percentages for P_{NTH} loads are presented in Figure 6.34, for two characteristic GSPs, i.e. GSP-14 and GSP-3, as in Section 6.3. The corresponding results for the thermal heating - P_{TH} loads are presented in Figure 6.35, for the same GSPs. All percentages are calculated using (6.8) and based on the Q_{min} reference power factor.

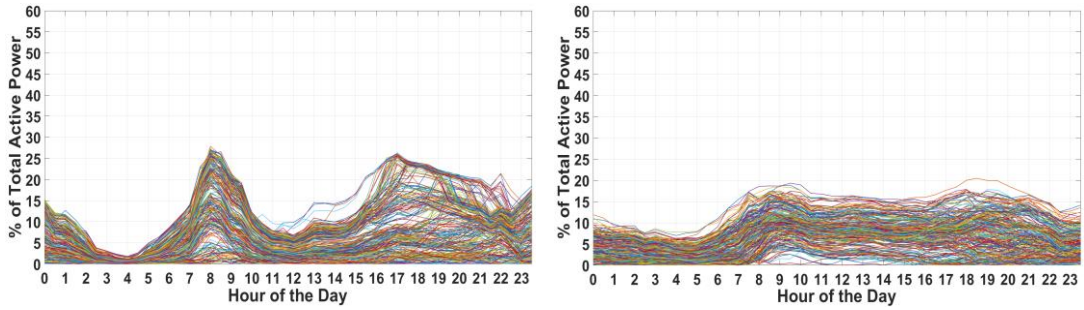


Figure 6.34: Estimated P_{NTH} loads for: a) GSP-14 and b) GSP-3

Estimated participation of P_{NTH} loads (Figure 6.34) peaks during the morning and afternoon to evening periods between 06:00 to 10:00 hours and 17:00 to 22:00 hours respectively, while participation of P_{TH} loads (Figure 6.35) is maximized during the night. In both cases, the results are more pronounced for the predominantly residential than for the commercial/mixture GSP. This is due to the higher percentage of seasonally variable loads for the residential GSPs, as presented in Section 6.1.5. Moreover, P_{TH} results, based on the current approach, are comparable with the results presented in Section 6.3 based on the multiple-regression disaggregation, in Figures 6.16 and 6.17.

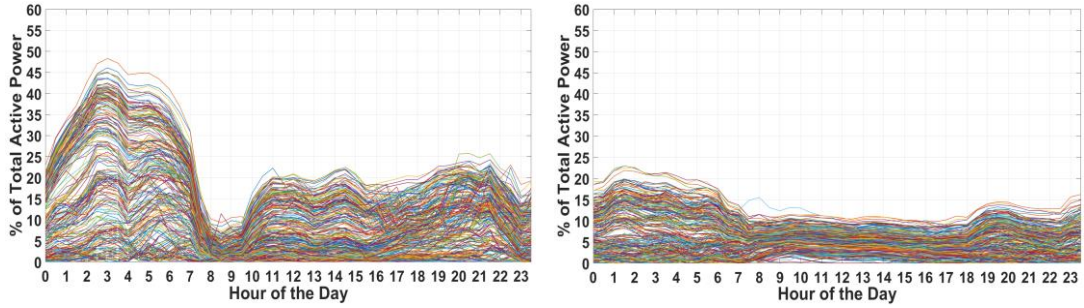


Figure 6.35: Estimated P_{TH} loads for: a) GSP-14 and b) GSP-3

The synoptic results from 77-GSPs, regarding the percentage contributions from P_{NTH} loads are presented in Figure 6.36 (a). P_{NTH} loads are shown to reach peak levels of ~ 15% during winter for the 95th percentile values with a median at ~8 % and the 5th percentile values at ~4 %. The seasonal minimum is found during mid-July when loads range between 0 to ~5% of total measured active power among all analysed GSPs. Figure 6.36 (b) shows the corresponding synoptic results for the P_{TH} loads. These have maximum values during winter between ~15 % to 30 % and ~20 % for the median values and minimums during mid-July, ranging between 0 % to ~5 %. Compared with the results presented in Section 6.3, the current methodology produces estimations that are approximately 5 % higher during the winter period, i.e. ~20 % for the current median value, compared with ~ 15 %, for both the *P-QT* and *P-QTE* models. Similarly, there is an approximately 5 % difference for the 5th as well as for the 95th

percentiles, with higher estimations shown, in each case, for the current methodology. The seasonality patterns and periods of maximum and minimum thermal-heating demands coincide for the two methods.

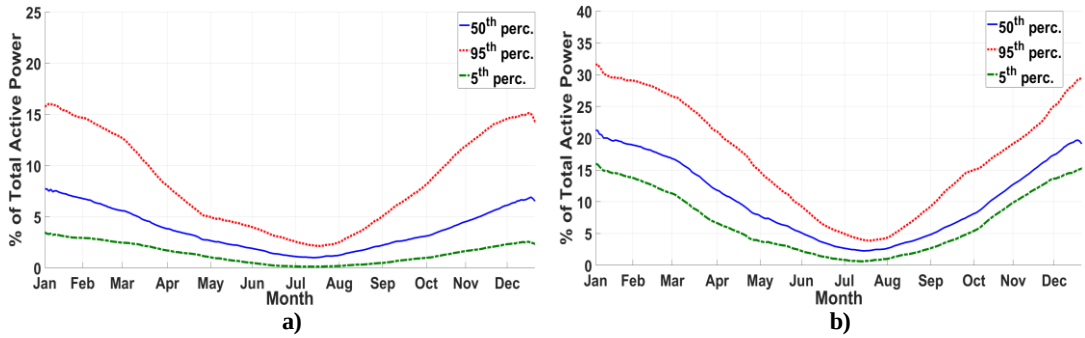


Figure 6.36: Estimated percentage contributions from: a) P_{NTH} and b) P_{TH} loads, for 77 GSPs

The results presented based on the power factor axis-transformation approach demonstrate that it is possible to decompose the variable portion of the load in portions that differentiate not only according to weather sensitivities but also according to (assumed) electrical characteristics. The distinctions made between P_{TH} and P_{NTH} loads show that there are regular/consistent seasonal changes in demands that are not only related to changes in thermal heating loads. The range of P_{NTH} loads is comparable with the results presented for lighting load disaggregation in Section 6.2, which indicates that both methods produce results that include demand variability that includes, but is not necessarily fully related to lighting loads. This is confirmed, for example in [106], where smart meter and survey based results show that seasonal variability exists for cold loads and wet loads, but to a lesser extent than for thermal and lighting loads. A reasonable assumption, based on the current analysis, is that such variability also exists for the use of electronics such as audio-visual and computer devices, related to variable occupancy levels and particularly when considering primarily residential GSPs.

6.5 Validation of Thermal-Heating Load Disaggregation

Precise and direct validation of the presented disaggregation methodologies would require knowledge of the specific disaggregated load contributions from the individual load categories discussed in the previous sections, which would also need to correspond to specific aggregation points, i.e. GSPs. Such data is, however, unavailable and therefore the validation approach presented in the current section relies on the approximate agreement between the IGZ consumption estimates, discussed in Chapter 5, Section 5.3.2 and the disaggregated loads of Sections 6.3.1 and 6.4. Validation is thus presented for the disaggregated thermal heating

load contributions, as these are the only loads for which validation data is directly accessible through the total economy-7 energy consumption statistics.

The procedure is based on the assumption that the total sum of the disaggregated thermal heating loads, i.e. P_{TH} (converted into units of energy - kWh), should be, approximately, equal to the portion of the economy-7 (E7) consumption which is of excess to the ordinary-residential (OR) consumption, as derived for the IGZ data and for each individual GSP under consideration. As the difference between the E7 and OR consumption is assumed to be directly related to night-hour economy-7 tariffs (used for space and water heating) the calculations are concentrated on the period between approximately 11:30 and 07:00 hours (performed in iteration-steps of increasing night period duration), which is denoted as *NH* in (6.17), i.e.:

$$No.E7(\overline{E7} - \overline{OR}) \approx \sum_{NH} P_{TH} \qquad (6.17)$$

where $No.E7$ corresponds to the number of economy-7 meters, at each GSP, $\overline{E7}$ and $\overline{OR}$ correspond to the mean economy-7 and ordinary-residential energy consumptions (annual), per meter (at each GSP), while the right-hand-side summation corresponds to the total energy consumption, during night-hours (i.e. *NH*), as estimated from the either the multiple-regression (Section 6.3.1) or the power-factor (Section 6.4) disaggregation methods (converted from units of power to units of energy, i.e. total annual consumption in kWh). The validation results are presented for the 11 GSPs for which IGZ data was retrieved, as discussed in Chapter 5.

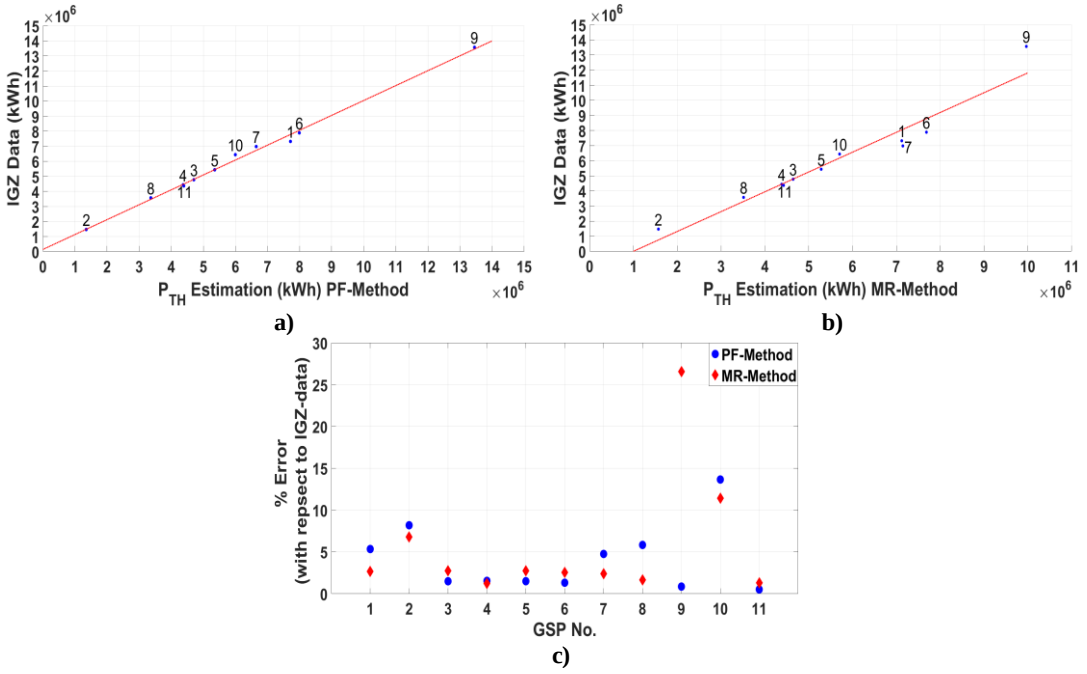


Figure 6.37: Scatter plots of IGZ consumption data compared to P_{TH} estimates form the PF-method in a) and the MR-method in b) and % error with respect to the IGZ data, for both methods, for 11 GSPs in c)

Figures 6.37 (a) and (b) are scatter plots of total annual energy consumption, where, in both plots, the y-axes correspond to the left-hand-side of (6.17), for each of the 11 GSPs and the x-axes correspond to the P_{TH} night-period energy consumption from the PF-method, in (a) and the MR-method in (b). Figure 6.37 (c) shows the percentage error between the two estimated consumptions, for both methods, with respect to the IGZ estimates, and for each GSP. The error is shown to be below 10 %, for both methods, for 9 out of 11 GSPs, with the highest errors found for GSPs 9 and 10 and, in particular, for the MR-method, reaching a maximum error of ~25 %.

Despite the fact that the results indicate good/satisfactory model performances (error below 10 %), there are a few important considerations regarding the validation methodology which should be discussed. Firstly, and as previously mentioned, the analysis is performed in iteration steps, where in each step the window of night-hour period is increased from 11:30-12:00 (i.e. smaller window of a 30-minute duration) up to 11:30-07:00 hours (i.e. largest widow of 7.5 hours' duration). In each iteration step the total energy consumption (as calculated from the RHS of (6.17)) is compared to the total IGZ energy consumption (as calculated from the LHS of (6.17)). The results presented in Figure 6.37 correspond, for each GSP, to the window length which minimises the error between the two thermal heating consumption estimates and it therefore represents the best possible agreement between disaggregated loads and validation data. These "error-minimisation" periods are, however, not equal in length among GSPs and particularly among the two disaggregation methods. On average, over all 11 GSPs, the error is minimised for a window length which spans from 11:30 to 04:00-05:00 hours, for the PF-method, while the period of error minimisation shifts to approximately 06:00-07:00 hours for the MR-method. This means that the PF-method (Section 6.4) produces higher estimates (and higher night-period estimates) for the total thermal demand, compared with the MR-method (Section 6.3.1). This differences have also been discussed with respect to the results presented in Figure 6.36, in the previous section. Since the extent of the night-period window is less than the corresponding duration of economy-7 tariffs (which, as the name suggests, covers a 7-hour period) for the PF-method, it can be assumed that the disaggregated loads may include non-thermal/heating demands, whereas, in the case of the MR-method, the window extends (for a number of GSPs) over the prescribed E7 tariff period, which may suggest an underestimation of the corresponding loads. However, it is also possible that requirements for heat storage and water heating applications are met before the completion of the 7-hour period, in which case the results would suggest a more accurate representation of the corresponding load categories from the PF-method. It should also be mentioned that, while the analysis presented in Sections 6.3 and 6.4 is demonstrated on weekdays only (and excluding Christmas holidays), the

validation results required the inclusion of all 365 days of the year. This, in turn, required a separate analysis of weekends and the inclusion of days with atypical consumption patterns (i.e. holidays) which may have potentially decreased the overall accuracy of the described methods.

Furthermore, although the two methods produce comparable results (which was also demonstrated for the average seasonal contributions of thermal heating loads in Figures 6.18 and 6.36), the two are not equivalent regarding the resulting estimations per GSP. A comparison is presented in Figure 6.38 based on the analysis of all GSPs for which active and reactive demand data were available. Figure 6.38 (a) shows the average thermal-heating loads considering all hours of the day, per GSP, for the PF-method (x-axis) and for the MR-method (y-axis), i.e. scatter-plots show average thermal loads for the two methods for each of the 77 GSPs used in the analysis. The correlations can be improved, as shown in Figure 6.38 (b), by considering the average load contributions for the night hours only, i.e. between 23:30 to 07:00 hours (since during this period the seasonal changes have been shown to be mostly associated with heating demands). The correlation coefficients are ~0.5 in (a) and ~0.7 in (b), indicating improved consistency between the two approaches for the night hours (and also more consistent results with respect to actual consumption).

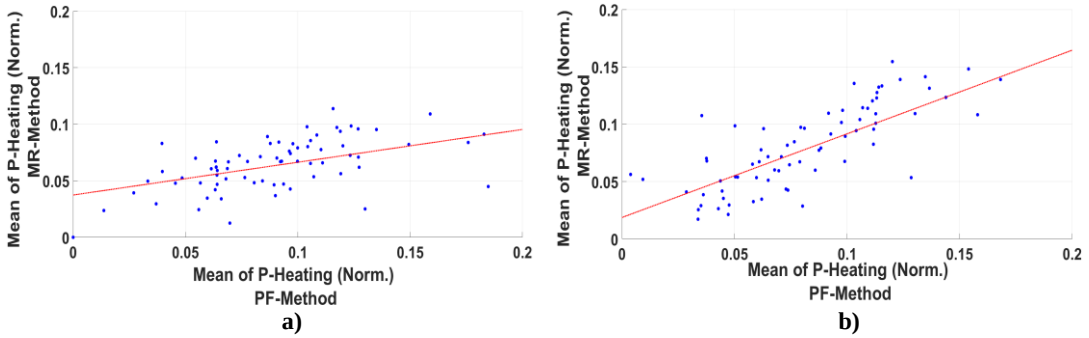


Figure 6.38: Mean normalised thermal-heating demand, per GSP, for the PF and MR methods: a) all hours of the day and b) night-hours only

Finally, it should be mentioned that the per half-hour correlations can be further improved, with correlation coefficients above 0.85, when the residuals are included in the resulting load estimations in the MR-method, i.e. by adding the residuals of the multiple regression surface fits to (6.7). This was a particularly interesting result which indicates that, although the two methods do not produce equal estimates, the inclusion of the residuals in the MR-method produces estimates that closely match the seasonal variabilities of the PF-method. Although these observations are not further investigated in the current work, it is speculated that the power-factor transformation methodology presented in Section 6.4, is, in fact, very closely

related to multiple-regression in the sense that it is based on similar error/residual minimisation criteria, but performed as an iterative optimisation process.

6.6 Chapter Conclusions

The initial decomposition approaches (Sections 6.1 and 6.2) showed that it is possible to make distinctions in the measured active power demand envelopes, based on the annual, daily and half-hourly base values. These distinctions can be used for determining the expected base, intermediate and peak loads in the system, as well as to define temperature and, to a lesser extent, solar irradiance balance-point values. Distinctions in active power levels can be used for generation and in particular distributed/renewable generation planning, as these are connected to the distribution networks and therefore demand expectations must be assessed for the corresponding locations and characteristic consumptions of connected customers. In this context, demand expectations can be estimated for smaller subsets of the annual data to represent the portions of total demand for specific periods of the year. In contrast with forecasting, peak demands, as presented in this chapter, do not correspond to single-value maximum demand expectations, but rather to portions (e.g. percentages) of the loads with specific characteristics, such as seasonal variability. Similarly, knowledge of the base and intermediate (and seasonally constant) portions of total demand can also be used to calculate consumption profiles that are more representative of the diurnal consumptions, i.e. excluding the influence of seasonal loads and therefore study the demands and, possibly, the efficiency of base-load appliances, e.g. cold loads and stand-by loads, or loads that vary through the diurnal cycle, e.g. consumer electronics, wet-loads, etc. These profiles can also be used to estimate the availability of shiftable-static-loads (SSL), discussed in Chapter 2.

Regarding thermal loads, the necessity for more sophisticated disaggregation approaches arises from the fact that the seasonal demand difference, i.e. measured demand minus balance-point demand, cannot be regarded as being used exclusively for heating/cooling purposes. In other words, the seasonally variable portion of active power is known to be comprised of several load categories. Accordingly, the multiple-regression (Section 6.3) and PF and data-transformation (Section 3.4) approaches have been developed, in order to account for the effects of multicollinearity between the available explanatory variables, as well as to capture loads with specific electrical characteristics (and based on necessary assumptions).

A proposed expansion for the presented methodologies is the inclusion of more detailed component based load-models, e.g. exponential, polynomial, etc. (discussed in Chapter 2), for selected load-categories and therefore modify the analysis to account for variations in voltage

levels. With the addition of load models, it is possible to implement disaggregation on the total measured demand values, as opposed to the variable portion, and therefore produce accurate estimates for a more diverse set of load-types. In the same context, the inclusion of more explanatory variables can potentially reduce some of the necessary assumption and help distinguish between load-types that, in the current approaches, are limited to being represented in common groups (e.g. thermal heating loads probably include percentages of water-heating loads).

Information on the relative contributions from disaggregated load-types can be used in subsequent studies for several purposes. Seasonally variable demands for particular loads are potentially more suitable as deferrable loads, as these are determined by sensitivities to external parameters and are not as closely linked to behavioural patterns as other load types are (and therefore can be more manageable). In the same context, thermal efficiency for selected locations can be studied through the heating/cooling disaggregation approaches and with additional socio-economic data, these can be used to assess the efficiency of electricity, gas or other fuel types for thermal energy. The per half-hour granularity of the analysis offers better resolution and allows to determine the concentration of the disaggregated loads for specific diurnal periods, thus enabling more specific DSM interventions, e.g. incentives for off-peak tariffs or for tele-switched loads can be formulated for specific locations and customer-classes based on the information for disaggregated loads. The results can also be used for the validation of bottom-up aggregation models, as these are usually compared to survey-based information or depend on large scale and costly data retrieval from metering devices at the individual household level (and SM-data limited by privacy related concerns). The current analysis can also provide information on the probability of occupancy, derived directly from the demand measurements, i.e. the per half-hour ratios of seasonal demands to seasonally-constant demands can be assumed to be related to occupancy levels.

Chapter 7: Load Forecasting

The forecasting methodology presented in this chapter is based on multiple-regression and follows from the results presented in Chapter 4, Section 4.7, which demonstrated that multiple-predictor variables can produce more accurate demand models, with higher coefficients of determination (R^2) and reduced seasonal autocorrelations of the residuals, compared to single-predictor variable models. The analysis presented in Chapter 4 also showed strong correlations between the electricity demands (primarily active power) and the analemma parameters, i.e. solar elevation and solar azimuth angles, and these are therefore included in the sets of predictor variables. The approach is "dynamic", in the sense that all possible pair-combinations of available predictors are considered and the final models are selected automatically according to their performance on the training datasets, which are comprised of data-subsets, i.e. days of the week and half-hours of the day, for which different demand patterns have been established in Chapter 3.

Section 7.1 discusses the models' specifications, the number of predictor-variable combinations, the separation of demand and predictors into matrices of half-hours and days-of-the-week and the active and reactive power training and validation datasets. Section 7.2 describes the model selection process, which is based on goodness-of-fit indices and Section 7.3 presents the results for the medium-term demand forecasting of active and reactive power, for individual GSPs, as well as for the distribution of errors according to the diurnal and weekly time frames. Finally, Section 7.4, describes a simple modification in terms of a residual correction-factor, that can be used to adjust the presented models for short-term load forecasting, demonstrating improved performance from one and up to several half-hours ahead.

7.1 Methodology

The analysis is concentrated on the Scottish-B dataset, i.e. seven GSPs as presented in Chapter 3, for which more than one year of measurements were available. The seven GSPs are presented in Table 7.1, according to their percentage-contributions from total residential and economy-7 residential demands, as estimated in Chapter 5 based on the customer-class disaggregation approach. The GSP numbers given in parentheses correspond to the overall data availability (i.e. out of 98 GSPs), but for convenience, GSP-numbers from 1 to 7 are used in this chapter. The percentages are provided in order to demonstrate that the forecasting methodology can be applied to GSPs of diverse customer-class mixtures, which range from

approximately 35 % to 90 % residential demands (thus 65 % to 10 % industrial and commercial demands), as well as from various mixtures of ordinary residential and economy-7 residential demands.

Table 7.1: Estimated % contributions from total residential (TR) and economy-7 (E7) demands

GSP No.	% TR	% E7
1 (47)	~80	~20
2 (48)	~80	~15
3 (49)	~80	~25
4 (50)	~95	~20
5 (51)	~60	~20
6 (52)	~50	~15
7 (53)	~40	~10

For active power, all possible combinations of two predictor variables are considered including: temperature -*T*, solar irradiance - *S*, solar elevation angle - *E* and solar azimuth angle - *A*. For reactive power, poor modelling accuracy and thus forecasting performance can be shown when considering only meteorological and analemma parameters and therefore active power - *P* is included in the set of predictors. The weak correlations between reactive power and meteorological/analemma parameters have been demonstrated in Chapter 4 and are also discussed in the context of modelling accuracies in Section 7.2.

Table 7.2: Model specifications (*P*-active power, *Q*-reactive power, *T*-temperature, *S*-solar irradiance, *E*-elevation angle, *A*-azimuth angle)

Pairs/Combinations of Two-Predictors: (x_A and x_B)		**Polynomial Degree: (n_A and n_B)**
for Active Power (P):	***for Reactive Power (Q):***	***applies to both:***
T & S; T & E; T & A S & E; S & A; E & A (6 combinations)	P & T; P & S; P & E; P & A; T & S; T & E; T & A; S & E; S & A; E & A; (10 combinations)	1^{st} & 1^{st}; 1^{st} & 2^{nd}; 1^{st} & 3^{rd} 2^{nd}& 1^{st}; 2^{nd}& 2^{nd}; 2^{nd}& 3^{rd} 3^{rd}& 1^{st}; 3^{rd}& 2^{nd}; 3^{rd}& 3^{rd} (9 combinations)

For both variables, model flexibility is further increased by the inclusion of 1^{st}, 2^{nd} and 3^{rd} degree polynomials for each *X* degree of freedom, i.e. for each predictor variable, and therefore the total number of combinations is given by:

$$\frac{n!}{k!(n-k)!} \times d^2 \qquad (7.1)$$

where *n* denotes the number of predictor variables, i.e. $n = 4$ for active power and $n = 5$ for reactive power, $k = 2$ for pairs of predictor variables (i.e. two per model) and $d = 3$ for the maximum allowed polynomial degree, giving a total of 54 and 90 models, for active and reactive power respectively. These specifications are summarised in Table 7.2.

All models are evaluated for the data-points corresponding to days of the week – *d*, at specific half-hours of the day – *t* and therefore dimensionality is increased so that active and reactive power demands per GSP are represented by matrices of size $d \times t = 336$, for $d = 1,..,7$ and $t = 1,..,48$. Each matrix element is a set of demand measurements at a particular *d* and *t*, thus including 52 data-points per calendar year. The same data preparation is applied to the corresponding measurements of the independent/predictor variables. This "separation" of measurements is implemented in order to increase forecasting performance and it is based on the results presented in Chapter 3, which showed significant differences in demand levels and in seasonal variability, for different half-hours of the day and different days of the week (more pronounced between weekdays and weekends). Similarly, and based on the results presented in Chapter 4, the sensitivity of the regression coefficients on *d* and *t* has also been shown, for the seasonal and seasonal per half-hour correlations between active/reactive power and meteorological and analemma parameters.

The fitted surfaces, with predictor variables x_A and x_B and polynomial degrees n_A and n_B are given, for each unique *d* and *t* combination, per GSP by:

$$y(x_A, x_B) = \sum_{i=0}^{n_A} \sum_{j=0}^{n_B} (\beta_{ij}\, x_A^i\, x_B^j\, k),$$

$$k = \begin{cases} 0, & i + j > z \\ 1, & otherwise \end{cases}, \; z = \begin{cases} n_A, & n_A > n_B \\ n_B, & otherwise \end{cases} \qquad (7.2)$$

where *y* is the estimated active (or reactive) power and β are the model coefficients. Note that the interaction terms between x_A and x_B cannot have a total polynomial degree greater than the maximum of set $\{n_A, n_B\}$, thus the constraint set by *k*. For example, for a model with $n_A = n_B = 3$, the β-coefficients will have indices: β_{00}, β_{10}, β_{01}, β_{20}, β_{02}, β_{30}, β_{03}, β_{11}, β_{12}, β_{21}, where β_{00} is the y-intercept.

The models described in Table 7.2, are evaluated for the all $d \times t$ datasets, thus giving a total of 18144 and 30240 multiple regression surface-fits per GSP, i.e. $54 \times 7 \times 48$ for active power and $90 \times 7 \times 48$ for reactive power. The multiple-regression models are estimated based on an OLS algorithm, as discussed in Chapter 4, Section 4.3. The computation time is approximately 30-minutes per GSP and based on processing performed on the PC-desktop described in Chapter 1.

Model performances are evaluated on the training datasets and are expressed in terms of R^2 values. The training datasets are constituted of demand measurements from 1st January 2007 to 31st December 2008. The selection process for the best-models, which are subsequently used for forecasting, is described in Section 7.2 and is based on an R^2 maximization criterion;

and the final set of best-models, per GSP and per dependent variable (for active and reactive power) is of dimensions $d \times t$. For comparison, the best-models including only analemma variables as predictors are also selected, but for active power forecasting only (low modelling performance in the case of reactive power). An important advantage of the analemma variables is that these can be accurately calculated for any period and geographical location, whereas meteorological parameters need to be forecasted themselves and therefore, in practice, forecasting performances can be compromised when meteorological inputs contain inaccurate predictions.

Forecasting is implemented for the period between 1st January 2009 to 31st December 2009 and the measured active and reactive power demands of the same period are used to evaluate and quantify the forecasting performances (this is discussed in Section 7.3). A description of the training and validation datasets is given in Table 7.3.

Table 7.3: Description of training and validation datasets

Training Data				**Validation Data**			
for P		***for Q***		***for P***		***for Q***	
X	***Dates***	***X***	***Dates***	***X***	***Dates***	***X***	***Dates***
T, S, E & A	01/01/2007 - 31/12/2008	P, T, S, E & A	01/01/2007 - 31/12/2008	T, S, E & A	01/01/2009 - 31/12/2009	P	01/01/2009 - 31/12/2009 (forecasted)
						T, S, E & A	01/01/2009 - 31/12/2009

The forecasting results are computed based on the actual/measured meteorological inputs, corresponding to the validation data period. This can be considered unrealistic since forecasted weather conditions will not always match actual weather, particularly in the long-term and thus the weather forecast errors will have an impact on the model performances. However, average temperatures and solar irradiance levels are not expected to largely deviate from the mean (moving-average), at least when considering average daily values throughout one calendar year. Alternatively, forecasted meteorological data or data corresponding to the training datasets could be used for the final evaluation. As previously mentioned, these considerations do not apply for solar elevation and solar azimuth angles.

For reactive power forecasting, the validation data is as described above, except when models include active power as an independent variable. In these cases, forecasted active power, which corresponds to the validation period is used, instead of actual active power measurements. This implies that active power is forecasted first, for the 2009 year, and the results are then used as inputs for reactive power forecasting.

While the initial Scottish-B dataset includes six years of available records, a smaller subset is chosen for model-training, forecasting and validation, as described above. This choice is based on an inspection of the consistency of demand measurements over the 6-year period, as well as on issues of computation time. Issues of consistency are mostly related to reactive power measurements, which show more erratic changes from one-year to the next, for a number of GSPs. Problems in reactive power modelling have also been discussed in Chapter 3, Section 3.3, i.e. Fourier signal reconstruction, and are also reflected in the model selection process in Section 7.2, as well as in the final results presented in Section 7.3. Inconsistencies due to trend-elements, i.e. regular increase or decrease of active/reactive power demands from one year to the next, have not been found within the 6-years of available measurements. Over-yearly trends have been shown to have significant impact in long-term forecasting studies (as discussed in Chapter 2), when social, economic and technological changes affect overall demand levels and in such cases, it is necessary that these parameters are taken into consideration. The models discussed in this section can therefore be regarded as effective for medium-term demand forecasting.

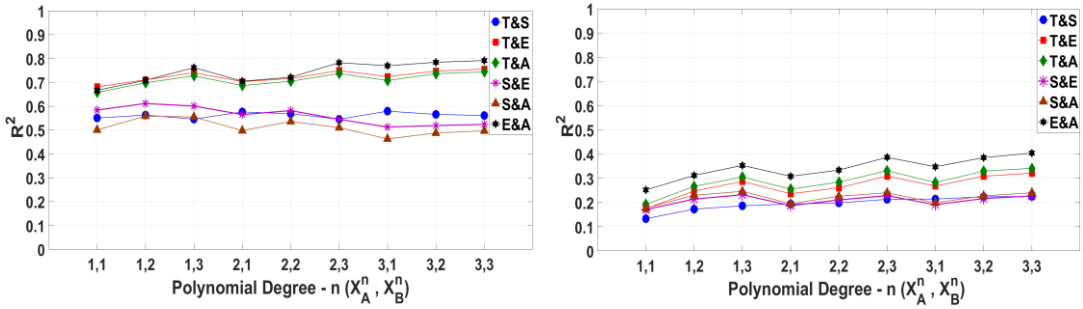
In Section 7.4, a modification is applied that enables short-term load forecasting, by the inclusion of a correction factor with which the predictions for the next time-steps are adjusted to new levels according to the residuals between actual and forecasted demands in the previous time-steps. This method is therefore suitable when real-time forecasting evaluation is possible and similarly, it can be expanded to include inputs from short-term weather forecasts.

7.2 Model Selection

According to (7.1), there are 54 models for active power and 90 models for reactive power (36 of which include active power as an additional predictor). The performance[22] of each of these models is evaluated, in terms of the coefficient of determination (R^2), for each unique combination of d and t (day of the week and half-hour of the day), thus producing $7 \times 48 = 336$ R^2 values per model, per GSP. Therefore, a summary of the performance of each model can be calculated based on the average R^2 value over all GSPs, i.e. average of $336 \times 7 = 2352$ values. These results are presented in Figure 7.1 (a) for active power and in Figure 7.1 (b) for reactive power. For reactive power, only the models based on the meteorological/analemma parameters are presented. In the case of active power, poorer overall performance is shown for

[22] Performance, in the current section, corresponds to the goodness-of-fit with respect to the training data and not to the forecasting performance as evaluated on the validation data, which is discussed in the next section.

the models including solar irradiance, primarily due to the absence of measurements during night-hours, as the results are averaged over all half-hours of the day. Best overall performance is shown for E&A models, i.e. solar elevation and solar azimuth angles as the two predictor variables and using 3rd degree polynomial surfaces (i.e. $n_A = n_B = 3$), with an average coefficient of determination of ~0.8. Second and third best model performances are shown for T&E models, i.e. temperature and solar elevation angle and T&A models, i.e. temperature and solar azimuth angle, where in each case, a tendency of increasing R^2 values is shown for increasing polynomials.

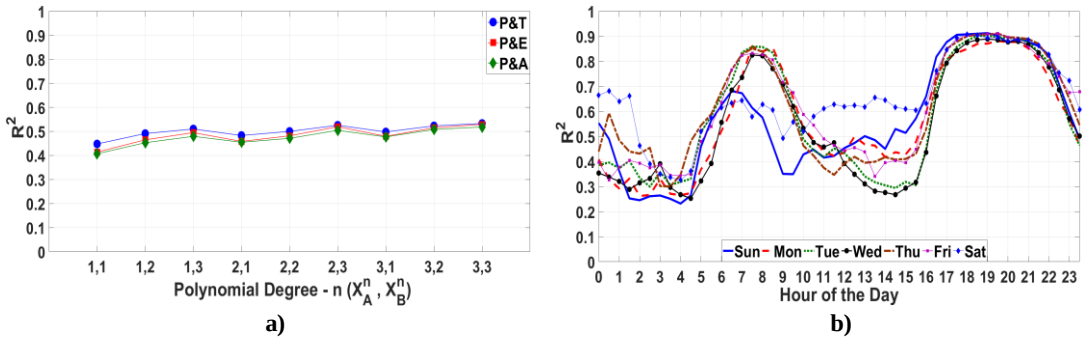


a) **b)**

Figure 7.1: Average model performances (over all half-hours, days and GSPs) for: a) active power and b) reactive power (excluding *P* from predictors)

For reactive power, poor performance can be shown for all models, with average R^2 values below ~0.4, in all cases. The general tendency of higher R^2 values for increasing polynomials is shown for reactive power as well, with E&A models of $n_A = n_B = 3$ maximizing the goodness-of -fit, which is however still below ~0.4.

Reactive power model performances can be improved by the inclusion of active power as one of the predictor variables, as discussed in Section 7.1. Marginally higher R^2 values are given for P&T, P&E and P&A models, but still below an average value of ~0.6, as shown in Figure 7.2 (a).



a) **b)**

Figure 7.2: a) average model performance (over all half-hours, days and GSPs) for reactive power (including *P* in predictors) and b) performance of best-models for reactive power, per day of the week and half-hour of the day

The poor performances are not, however, homogeneously distributed among different days of the week and diurnal periods, as it is shown for GSP-1 (47), for the R^2 values of all selected best-models, per *d* and *t*. Despite the fact that the goodness-of-fit is above ~0.8 for certain periods of the day, e.g. morning and evening hours for weekdays, overall forecasting performance for reactive power is expected to be low, or at least lower than for active power, for all GSPs.

While Figures 7.1 and 7.2 indicate the overall/average performance of each model, the results vary according to the specific *d* and *t* combination, e.g. the best model for 12:00 hours on a Sunday might be different form the best model at 23:00 hours on a Monday and, similarly, the best models can vary among GSPs. Accordingly, a unique set of models (i.e. best-models) is selected for each GSP, where each set is comprised of $d \times t = 336$ models, where each element of the set corresponds to the model with the highest R^2 value among the 54 tested for active power and the 90 tested for reactive power (and therefore each set is not made-up from 336 unique models). For active power, the best-models are, primarily, of the 3rd polynomial degree (i.e. $n_A = n_B = 3$) and more specifically, considering all seven GSPs: E&A model (~55 %), T&E model (~25 %) and T&A model (~15 %), given as percentages over all selected best models from all GSPs. Similarly, for reactive power, the sets of best models are comprised of: P&T model (~40 %), P&E model (~30 %) and P&A model (~18 %).

Considering the selected terminology, it should be noted that the use of the phrase *best-models* is based on the model performances prior to the forecasting analysis. Conversely, *optimal-models* is used to denote the best models, based on their performance on the validation datasets. These correspond to the models that give the lowest forecasting error, but which are not necessarily the ones that satisfied the R^2 maximization criterion as discussed in the current section, i.e. they can only be determined post-validation and are therefore discussed in the following section.

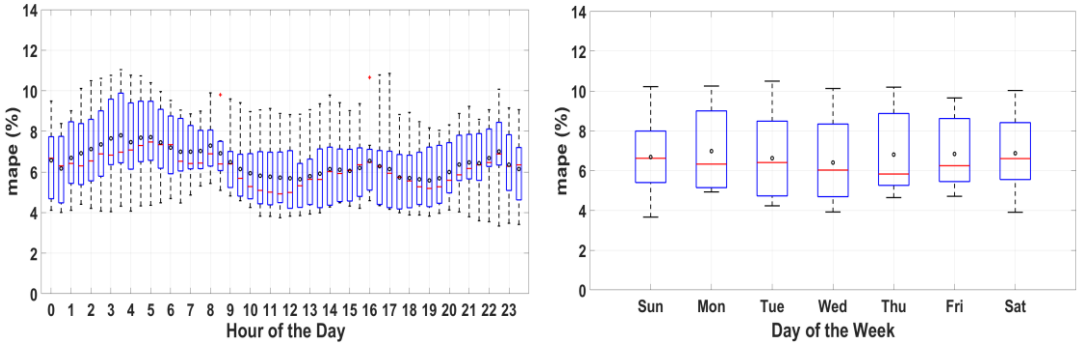
7.3 Forecasting Performance

Active and reactive power demands are forecasted for the validation year, form 1st January 2009 to 31st December 2009, using the corresponding pairs of x_A and x_B predictor variables as inputs to the best-models specified by the R^2 maximization criterion, discussed in Section 7.2. Forecasting performance is quantified using the mean absolute percentage error, or mape, given by:

$$mape(\%) = 100 \times \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - y_i}{A_i} \right| \qquad (7.3)$$

where n is the total number of observations, A_i is the actual/measured active (or reactive) power and y_i is the forecasted value i, where $i = 1,2 \ldots, 17520$ for the total number of half-hour measurements within the one calendar year period. The forecasting performance can also be evaluated per day of the week, per half-hour of the day, or seasonally, by evaluating (7.3) for the selected subsets of the validation dataset.

Figure 7.3 presents the mape values for active power, for the 48 half-hourly periods in (a) and for the seven days of the week in (b). The range of mape is represented by the extent of whiskers, extent of boxes shows 25th and 75th percentile values, lines in boxes indicate the median values (50th percentile) and circles in boxes indicate the mean values. These results are for the average mape over the seven GSPs used in the analysis.



a) **b)**

Figure 7.3: Average mape (%) from seven GSPs per: a) half-hour of the day and b) day of the week, for active power forecasting

Overall performance is between 3 % to 11 % error and generally below 7 % for the mean and median values. Slightly poorer performance is shown for the night, i.e. for the period between 01:00 to 06:00 hours, while no significant performance differences can be reported between the days of the week, as shown in Figure 7.3 (b), neither for the average results nor for the results from individual GSPs.

Figure 7.4 shows the resulting mape per GSP, for active power and for best-model selection according to the R^2 maximization criterion (Section 7.2), as well as for the same criterion but restricted to the solar azimuth and solar elevation angle models (E&A), i.e. excluding models with meteorological parameters. Figure 7.4 also shows the mape performance for the optimal-models, which is the potential performance assuming that the models with the lowest mape can be selected post-validation. This is therefore the resulting forecasting performance per GSP when mape is calculated for all available models (54 for active power from Table 7.2) and selection is performed afterwards, according to the best results per d and t. As a consequence, these combination of models shows better performance for all GSPs, but only marginally, with a maximum deviation for GSP-2 at approximately 1 % from the actual

forecasts. The results clearly indicate that performance of the models including only analemma variables is close to that of the combination of best-models, with noticeably poorer performance only for GSP-6. In two out of seven cases (GSPs 5 and 7), the analemma variables performed marginally better.

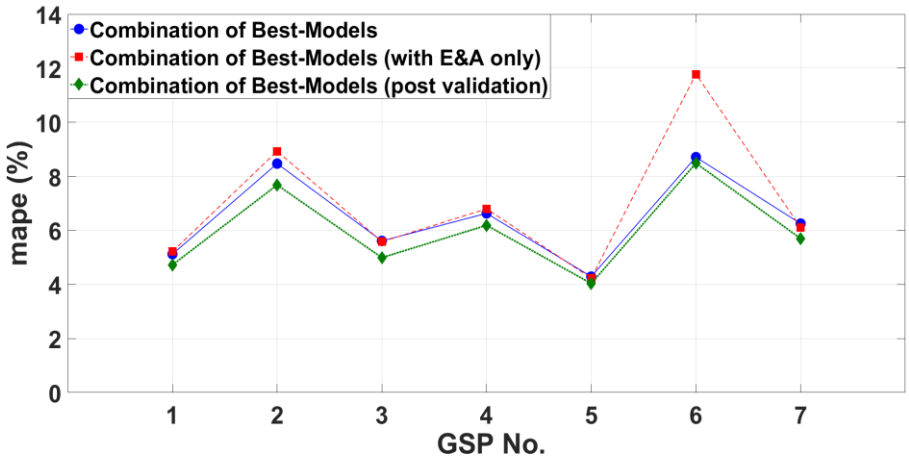


Figure 7.4: Resulting mape (%) per GSP, for active power forecasting

Since the analemma parameters are not subject to atmospheric phenomena and they are constant markers of diurnal and seasonal cycles they are more stable than the meteorological parameters in terms of the extent of variations (day-to-day fluctuations). This can be a disadvantage in cases when weather changes, such as a decrease in temperature levels, are associated with subsequent changes in active power demand, but as the extreme conditions are rare and demand does not always adjust to the extremes (e.g. saturation of demand response at very low temperatures, as discussed in Chapter 4), the analemma variables can perform better, on average. Moreover, as demonstrated by the post-validation model selection, the combination of models selected according to the maximum R^2 values is not necessarily the combination of models that can perform better in the validation phase and therefore while the initial selection (Section 7.2) produces sets of models primarily comprised of 3rd polynomial E&A, T&E and T&A combinations, the post-validation selection includes a wider variety of best models. Examples include: T&E with $n_A = n_B = 1$, T&E with $n_A = 1$ and $n_B = 3$, E&A with $n_A = 3$ and $n_B = 1$ and E&A with $n_A = 3$ and $n_B = 2$.

Figure 7.5 shows an example of active power demand forecasting for a one-week period during winter, for GSP-1. Actual measured values are shown together with forecasted values based on the combination of best-models, the combination of best-models with elevation and azimuth angles only and the combination of optimal models as selected post-validation.

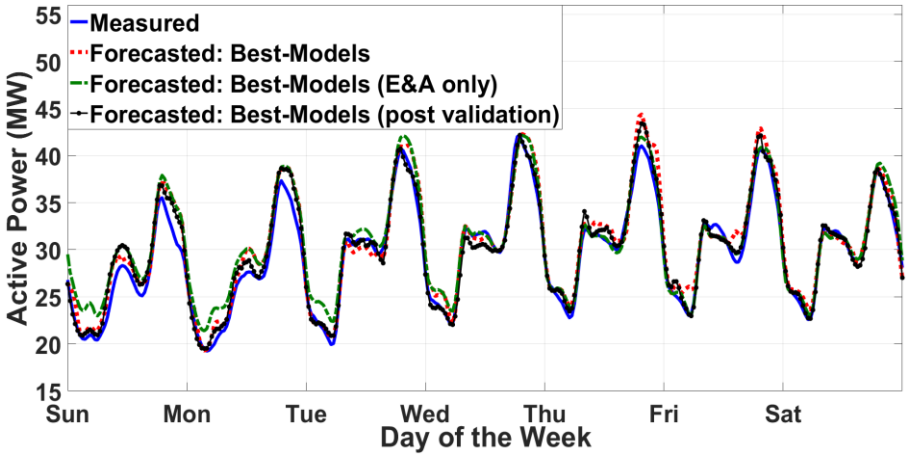


Figure 7.5: Example of forecasted demand, for GSP-1 (one week), for active power

Reactive power forecasting can be considered unsuccessful, even with the inclusion of active power demand in the models, with mape values generally above a 10 % error and very high inaccuracies for GSP-2 (~80 %) and GSP-3 (~40 %), as presented in Figure 7.6. The particularly high errors shown for these GSPs are attributed to their actual/measured reactive power demands. These include a high percentage of zero measurements during night hours, for GSP-2 and differences in demand levels between the training and validation datasets in the case of GSP-3.

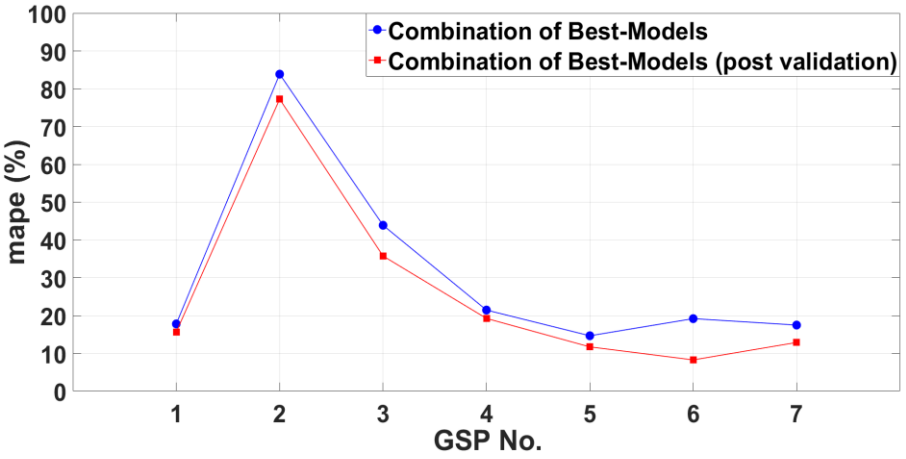


Figure 7.6: Resulting mape (%) per GSP, for reactive power forecasting

When excluding GSPs 2 and 3 from the final performance evaluation, the mean absolute percentage error is still moderate-to-high and within 10 % to 20 %. Mape values are also moderate-to-high for the post-validation model selection, indicating a general failure of the methodology for accurate sub-daily resolution forecasting for reactive power.

7.4 Correction Factor and Short-Term Forecasting

Forecasting performances can be considerably improved for half-hour ahead predictions by the addition of a correction factor for each forecasted value, in the form of the residual/error of the forecasted demand of the previous time-step, such that:

$$y_j^A = y_j + \varepsilon_{j-1} \tag{7.4}$$

where y_j is the initial forecast, as estimated in Section 7.3, ε_{j-1} is the residual between the initial forecast and the actual demand (of the previous time-step) and y_j^A is the new/adjusted forecast at time j, for $j = 1,2, \ldots, 17520$, for all half-hours the year (365 days).

The approach can be modified to account for more than one half-hour ahead predictions such that ε_{j-1} is replaced by ε_{j-i}, where i is allowed to vary accordingly. The results, for active power and reactive power, for up to 10 half-hours ahead, are presented in Figure 7.7 (a) and (b). In both cases, the original/initial forecasts were estimated based on the combination of best models, as described in Section 7.2.

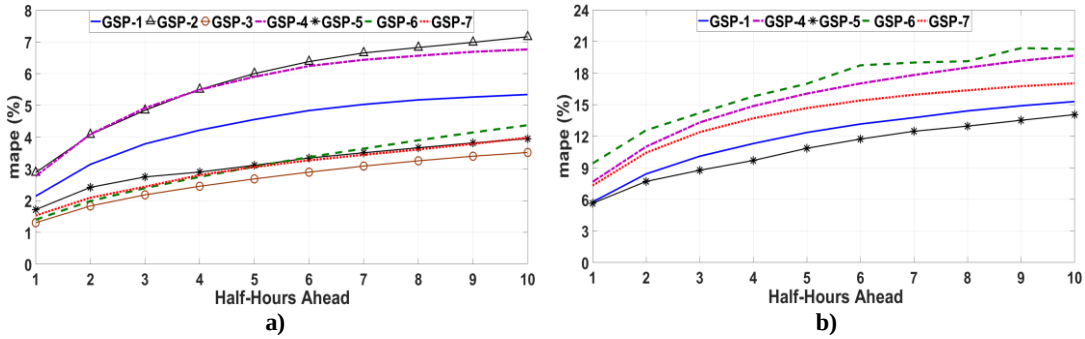


Figure 7.7: Short-term forecast results for: a) active power and b) reactive power (excluding GSPs 2&3)

The approach is of course restricted to when actual demand measurements from the previous half-hours are available, in order to be able to estimate the corresponding residuals. For active power, the error drops to less than 3 % for half-hour ahead forecasts, which is for all GSPs, at least as twice as good as the forecasts presented in the previous section. The errors steadily increase, but for up to five-hour ahead predictions, these are still below the levels presented in Section 7.3, apart from GSP-4. For reactive power, Figure 7.7 (b), GSPs 2 and 3 are not presented, following the discussion provided for the results in Figure 7.6. The error for the remaining GSPs drops below 10 % for half-hour ahead predictions and it converges to the levels of the medium-term forecasts, for more than five-hours ahead predictions.

7.5 Chapter Conclusions

The analysis showed that by not restricting the forecasting models to a specific set of predictors, higher flexibility is achieved and different inputs are selected according to the day of the week and half-hour of the day. These combinations are therefore more appropriate to the demand characteristics at the corresponding time-frames and for the consumption patterns of individual GSPs. Training of the models requires, however, at least one-years' worth of demand measurements.

The results also demonstrated that, in the case of active power, satisfactory medium-term forecasts can be achieved by relying on the analemma variables alone. The average performance, for one year's forecasts, was below 10 % error and only marginally worse than the performance of the optimal-models selected post-validation. This indicates that the selection process and the use of a simple goodness-of-fit maximisation criterion in Section 7.2 was successful. The introduction of a correction factor, i.e. residual of previous time-step(s) in Section 7.4, demonstrated short-term forecasting performances with less than 3 % mape, for all GSPs. This is of course provided that there is available real-time, or close to real-time forecast evaluation, while the performances are below the medium-term levels, for up to five-hours-ahead predictions.

Significantly lower forecasting performances have been shown for reactive power demands. These results are however in agreement with the analysis presented in Chapters 3 and 4, and support the previous conclusions about the reactive power irregularities and weaker dependency on external parameters. Medium-term forecasting performance is shown to be below 20 % (excluding GSPs 2 and 3), while the correction factor adjustment in Section 7.4, results in half-hour ahead prediction with mape below 10 %, which can (marginally) be considered within acceptable limits.

Chapter 8: Thesis Conclusions

This thesis has investigated the proposition that aggregate demands, measured at MV-network buses, contain not only quantitative information, but also substantial, qualitative and useful information on the specific characteristics of demand variability, evolution and composition. Accordingly, the presented work has demonstrated that, at least to a certain extent, this "useful information" is accessible through appropriate statistical analysis. In particular, the methodologies presented have been developed/adapted and applied for the purposes of: load-profiling and decomposition; determining the effects of medium/short term drives of demand variability; quantifying the statistical associations between electrical parameters; customer-class and load-type disaggregation; and medium/short term load forecasting. The main intention has been to present a top-down approach, starting with a very limited number of prior assumptions, where each chapter in the thesis builds on the results and conclusions of the preceding chapters, therefore increasing the level of detail while expanding the scope of the analysis.

8.1 Thesis Synopsis and Implications

After two introductory and background chapters (Chapter 1 and Chapter 2), measurements of active power, reactive power and voltage have been decomposed according to the most significant temporal-modes of variability, in **Chapter 3**. These have been shown to be primarily comprised of the seasonal (i.e. yearly and sub-yearly), the weekly and the daily (i.e. diurnal) cycles and the contributions of these cycles to the total range of variations have been quantified, for individual substations, as well as for the complete dataset used for the analysis. This resulted in two main conclusions. Firstly, the significance and order/rank of significance of these components is different for different electrical parameters (i.e. P, Q and V). Secondly, the periodicities of the same parameter are also diverse among GSPs, but not ambiguous, i.e. GSPs fall within a limited number of groups with respect to the modes of variability. The first result has implications regarding the factors determining P, Q and V variability, as well as their inter-dependencies, which is also related to load composition. The second result indicates that variability is determined by specific demand characteristics, e.g. contributions from different customer-classes, and that these results can therefore be used for further load modelling and load disaggregation studies.

For active power, daily and yearly cycles are the most significant modes of variability, responsible for approximately (on average), 40 % and 33 % of the total range of variations,

respectively. "Reconstructing" active power based on a small number of Fourier components performed particularly well, demonstrating regularity in *P* variability and lower penetration of, statistically, stochastic components. Reactive power variability is shown to be determined primarily by the daily component, at approximately 37 %, and to a lesser extent by the weekly and yearly cycles, at approximately 17 % and 13 %, respectively. Here, reconstruction based on a limited number of Fourier components performed poorer, indicating decreased predictability of the variations and higher irregularity in the reactive power demand patterns. The fact that the same periodicities are not equally responsible for both active and reactive power variability, has implications about load composition (i.e. changes in the contributions from different load types are more significant seasonally, rather than daily – also shown from the correlational analysis in Chapter 4). Regarding voltage, grid-code regulations restrict variability within limits of $\pm 6\,\%$ of the nominal. The analysis has shown that these fluctuations can be considered to be, mostly, stochastic (from a statistical perspective), a fact that is reflected in the poor modelling performance of voltage signal reconstruction from the first 100 Fourier components. Nevertheless, subsequent analysis has shown clearer diurnal and seasonal voltage level distributions, but for smoothed and/or averaged values (i.e. voltage fluctuations at the distribution level of the grid, which correspond to aggregate *P* and *Q* demands, have been shown to be, mostly, unpredictable).

Apart from demand modelling possibilities, the results from the Fourier analysis can also inform decisions for the implementation of demand-side interventions. For example, the variable portion of the load (from some selected base value), was shown to be distributed throughout the year and, to a lesser extent, from day-to-day cycles, for consumption in the residential load sector, while the reverse was shown for primarily non-residential load sectors (Chapters 3 and 5). These portions of the total demand have been shown to be associated with particular load categories and, as a result, targeting of specific deferrable (demand-manageable) loads will benefits from the identified distributions with respect to different time-scales.

Significant differences in *P*, *Q* and *V* levels in the considered datasets have been found with respect to different days of the week. Most prominent variability was, as expected, between weekdays and weekends, but different levels were also shown for Fridays and, to a lesser extent, for Mondays. The weekday/weekend distinction was shown to be sensitive to load composition with respect to customer-classes (i.e. load sectors), with higher *P* levels for residential consumption during weekends. Higher voltage levels have been shown for weekends (in contrast with the aggregate *P* and *Q* tendencies), which also coincide with the

maximum deviations from the nominal voltage and can, therefore, inform power studies related to optimal voltage control and, possibly, voltage stability. The differences in demand levels between weekdays and weekends have also been presented with respect to the diurnal time-frame, which allowed to identify the particular daily periods when load composition is primarily affected by socio-behavioural patterns (e.g. working schedules), as opposed to periods more sensitive to seasonal and therefore weather related loads (also discussed in the context of correlations and dependencies).

Aggregate demand and voltage profiles have been presented in the diurnal and seasonal perspectives and levels of similar and dissimilar consumptions have been identified. For the same time-scales, the probability of occurrence of maximum and minimum demands has been quantified, indicating concentration of maximum and minimum system loadings that varies according to customer-class composition, as well as throughout the year, i.e. occurrence of peak demands was shown to shift from afternoon to evening hours, according to a seasonal cycle. Although not explicitly demonstrated, these changes are shown to be associated with changes in weather conditions (primarily solar irradiance levels). Despite the differences in demand patterns between weekdays and weekends, the periods of maximum and minimum demands have been shown to be almost identical between the two groups, which indicates that peak-shifting DSM interventions can be applied without distinctions between the two groups. However, and as previously mentioned, the actual timing of these periods varies according to the seasonal perspective. As the analysis was based on demands measured in four European distribution networks (with the biggest portion of the dataset corresponding to DNOs operating in the UK), including substations of various customer-class mixtures, the resulting profiles can be regarded as descriptive of demand variability for networks at similar geographical locations and with similar climate characteristics (this also applies to the results from the analysis of dependencies to meteorological conditions).

Combined diurnal-seasonal demand profiles have also been presented with respect to the rate-of-change of active power and reactive power, clearly capturing the periods of maximum changes in the system loading conditions. This analysis also identified the variable portion that correlates with seasonal changes in weather parameters, but remains relatively constant for other periods of the day (i.e. the seasonal shift in the rate of change of demands corresponds to the afternoon and evening periods, while the rate of change remains constant for the morning period). This section of the thesis also postulated a connection between the higher rates of change of demands and the higher probabilities of occurrence of network faults (resulting in short and long supply interruptions), although this has not been investigated in depth.

Chapter 4 presented results based on correlation and regression analysis between parameters *P*, *Q* and *V*, as well as between these three parameters and meteorological and analemma variables. In the diurnal time-frame, correlations between active power and reactive power have been shown to be particularly strong for the majority of GSPs (for which the corresponding *P-Q* data were available). In contrast, weaker correlations between the two variables have been demonstrated in the seasonal analysis, as well as in the seasonal analysis for particular daily periods, indicating more homogeneous load composition changes, with respect to *P* and *Q* demands, within single days, than throughout the year. This was also illustrated by the power factor variability over the two time-frames and it is also supported by the results of the Fourier analysis in Chapter 3. Seasonal *P-Q* correlations are significant only at particular periods of the day, which are mostly concentrated in the morning and afternoon to evening hours, for residential demands, and during mid-day for non-residential demands, which is, again, a useful indicator for both disaggregation and for DSM implementation, as it allows to distinguish specific load compositions, according to both load-types and customer-sectors.

Diurnal correlations of demands with meteorological and analemma variables have been shown to be consequential of the diurnal cycle (i.e. day/night alternation), indicating that, within each day, electricity consumption is determined primarily by relatively stable socio-behavioural patterns and that the changes due to external conditions are detectable only with respect to the yearly cycles. In the seasonal time-scale, active power has been shown to be strongly correlated with temperature and solar irradiance. The strength of these correlations has been shown to be associated with load composition related to different customer-classes, indicated by the increased contributions from the residential-sector, where active power demands correlate strongly with the seasonal variability of temperature. This showed that commercial and/or industrial consumption is to a lesser extent affected by the changes in weather conditions. Seasonal regression analysis has also demonstrated strong active power correlations with the analemma variables (i.e. solar azimuth and solar elevation angles) and, as a result, these have been used in the subsequent analysis aimed at load-disaggregation and load-forecasting. Reactive power showed to be moderately-to-weakly correlated with meteorological and analemma parameters, indicating that reactive power requirements are primarily determined by load-categories that are to a lesser extent affected by weather conditions (compared to the seasonal variability of active power demand).

Correlation and regression analysis has shown no statistically significant associations between voltage fluctuations and active or reactive power demands, as well as between voltage and

external meteorological parameters, apart when filtered (i.e. smoothed) voltage values were used in the seasonal time-scale. However, and as shown by the Fourier decomposition, the seasonal and day-to-day voltage fluctuations are at comparable levels, which means that smoothing has the effect of over-fitting and therefore it does not reveal any meaningful statistical associations between the variables (although, and as mentioned in the Chapter 3, averaged/smoothed voltage levels reveal general tendencies of increased/decreased voltages through the diurnal and yearly cycles).

The sensitivity of demands to changes in weather conditions throughout the year has also been investigated, based on a moving-window regression approach. These results indicate the periods of maximised rate of change in electricity demand, as a result or coinciding with the corresponding changes in temperature and solar irradiance (or solar elevation angle), which are concentrated during the autumn and spring months. It was also shown that these changes could not be accounted for by simple linear regression models and require extensions to multiple-regression analysis for improved modelling performances, indicated by the extent and distribution of residuals between the two approaches. In the same context, it has been demonstrated that modelling limitations are not a result of non-linear relationships, even though these are not absent and are expected to be more important when analysing data from different geographical locations. Although not explicitly demonstrated, these results also suggest that some portion of the differences in electricity consumption during periods of similar temperature levels are, potentially, associated with psychological phenomena resulting from the passage from winter-to-summer months and, similarly, from summer-to-winter months. This observation may also be related to differences in solar irradiance levels, which can affect the perception of ambient temperature (apart from direct effects on the demand for artificial lighting). The assumption is that the proliferation of automatically switched or DSM-controlled heating/lighting equipment will result in a general decrease of such effects. Again, from the point of view of DSM and regardless of the actual cause, these results provide important additional information for the implementation of temperature-based DSM schemes.

Chapter 5 presented a classification of MV-substations based on the clustering of mean diurnal demand profiles. This was expanded into a more detailed customer-class disaggregation model, using metrics that quantified daily, as well as seasonal active power demand variability, where the results were presented according to four different customer-classes. This part of the analysis has demonstrated that load composition with respect to contributions from various customer-classes is a determining factor of daily and seasonal

demand profiles and therefore, identifying the variability of these patterns among different GSPs was sufficient for their classification.

Comparison of the clustering-based classification and customer-class disaggregation results has shown that simple clustering is both less effective and less accurate in predicting contributions from the various classes, and that more detailed feature extraction is required for that application. This indicates that simple assumptions regarding specific patterns and customer-classes (e.g. mid-day peak for the commercial sector), are inadequate for assessing contributions when dealing with aggregate measurements that are, in the majority of cases, mixtures of different percentages of various customer-classes. Examples of initial results have also been presented, demonstrating the possibility of modelling demand profiles according to selected percentages from different customer-classes and based on the relationships between particular demand characteristics and specific sectors, as established from the disaggregation method.

The methodology has possible applications that include: the validation and/or update of existing load/demand profiles that are used in tariff formulations, generating synthetic profiles that can be used for load-modelling studies, for the evaluations of demand-side interventions (i.e. different peak periods for different percentages of residential and non-residential loads), for the assessment of proposed DG-integration and for planning considerations. Importantly, assuming that a sufficiently diverse dataset is developed, the presented analysis relies on direct demand/consumption characteristics and not on survey based and socio-economic estimates (or the necessity for extensive LV-monitoring) and therefore the resulting models are more accurately representing specific locations, DNOs and the corresponding customer-mixtures under consideration.

Chapter 6 presented approaches for the decomposition of electricity demand according to different "levels", determined by constant and variable portions of measured active power. It was demonstrated that these distinctions can be used to define base, intermediate and peak portions of the total active power demand for the whole year, an approach which can be easily expanded to a short-term analysis and provide estimations for peak load requirements that can facilitate generation planning and, particularly DG, regarding specific locations, customers and GSPs. In this context, the peak load does not refer to a single-value estimation of maximum expected demand, but rather to the peak portion (e.g. percentage) of the load, that is distinguishable according to specific characteristics (e.g. seasonal variability). Similarly, the intermediate portion of the total load can be defined as seasonally constant, but variable within the diurnal cycle and it is therefore related to socio-behavioural consumption habits, specific

to customer-class mixtures. The remaining, which has been defined as the base, corresponds to the minimum system/GSP requirements and it is therefore representative of seasonally and diurnally constant portions of the load (thus related to e.g. stand-by electronics, etc.)

Two different approaches have been presented for the disaggregation of the total demand into generic and specific load-categories, primarily concentrating on thermal heating, thermal cooling and lighting loads, as well as on non-thermal loads that exhibit seasonal variability, such as loads related to occupancy levels (e.g. use of electronic devices). The disaggregation results can have various applications. As mentioned, various studies, as well as network operators, rely on survey-based estimations of contributions from different load-categories. The presented approaches however, can be used to determine these loads directly from the demand measurements. The results can inform DSM interventions, through determining deferrable portions of the load, or for studies of the efficiency of thermal load consumption, through the disaggregation of total demand into heating/cooling components, as well as for balancing variable demand and generation from renewable/DG sources. This chapter also presented results for base (i.e. threshold or balance-point) temperature and solar irradiance levels, which are specific to the UK (and particularly to the north of England and south/central Scotland) and are flexible up to a half-hourly resolution.

In **Chapter 7**, meteorological and analemma variables were used as predictor-variables in a multiple-regression based approach, for forecasting active and reactive power demands. The modelling performance of all possible pair-combinations of these predictors represented with both linear and polynomial fitting functions has been evaluated and the best combinations have been used for medium-term and short-term electricity demand forecasting. This analysis has revealed that the analemma variables performed only marginally worse (or in some cases equally well) than combinations of predictors including meteorological parameters. Effectively, this allows to develop regression-based forecasting models in the full absence of available historical weather data. The analysis has also shown that the models performed poorer for reactive power forecasting, a result that was expected based on the conclusions of Chapters 3 and 4, which indicated reduced periodicity-based reconstruction performance for reactive power (Fourier analysis), as well as weaker correlations between reactive power and meteorological and analemma variables. Furthermore, a simple modification in the form of a residual "correction-factor", has allowed short-term forecasting with performances of less than 3 % (mean-absolute-percentage-error), and improved results, compared to the medium-term analysis, for up to 5 hours-ahead predictions.

8.2 Research Limitations

In the customer-class disaggregation methodology, described in Chapter 5, the target-dataset has been limited to 11-substations, which correspond to a single DNO operating at particular geographical location. This limitation resulted in no available data to be used exclusively for validation, implying that the results are, potentially, more appropriate for the specific grid and consumption characteristics of the Scottish-GSPs (although the methodology included an analysis of the consistency of the resulting estimations, when applied to all GSPs). Furthermore, the format of the available consumption statistics did not include any distinctions between industrial and commercial (I&C) demands and, as a result, the corresponding estimations are with respect to a common I&C category (although the analysis showed that substations that are primarily-industrial could be identified as "outliers"). A more diverse sample of GSPs with known customer-class mixtures would have allowed for a more flexible disaggregation approach, as the methodology itself has been shown to perform well and produce consistent results among GSPs of different demand characteristics.

Similarly, the disaggregation results presented in Chapter 6 could not be explicitly validated, as these contributions are not available for the corresponding GSPs (apart from the approximate validation data, based on the IGZ consumption statistics, which was presented in Section 6.5). In the same context, while the presented thermal-heating disaggregation methods demonstrated similar (but not identical) results, in terms of seasonal percentage-contributions; differences were also shown for the estimations for individual GSPs. Which of these approaches performed better and what additional load-types can be identified based on their differences is, therefore, an open question. The presented methodologies were also implemented based on some necessary assumptions, e.g. relating seasonal resistive loads to electrical heating elements, where in reality, smaller percentages from seasonally variable water-heating, cooking, wet-loads (heating-elements), etc. cannot be explicitly shown to be excluded from the final estimations. Similar assumptions have been made for the correlations of solar irradiance/elevation angle levels and electrical lighting. It should also be noted that voltage measurements were not acquired until the latter stages of this PhD study and that the methodologies could have included voltage variability, which can be used to account for more specific load-categories, i.e. by the inclusion of component-based load models in the statistical analysis, as targets for the disaggregation. This can, potentially, render some of the assumptions mentioned above as unnecessary, as more detailed load models can differentiate between load-types that were included in common categories.

Finally, in the medium-term forecasting approach, actual meteorological measurements were used for the implementation of the forecasting models, i.e. actual weather measurements were used in the validation datasets. This can be considered unrealistic, since for practical applications the weather parameters need to be forecasted themselves and this will have an impact on the forecasting performances. Nevertheless, the analysis has shown that the models performed marginally worse, or equally well in some cases, when the analemma variables were used as the only predictors (thus minimising the need for accurate weather forecasts).

8.3 Further Work

- The analysis of electricity demand measurements based on discreet Fourier transforms has shown that this can be expanded into a more comprehensive and detailed modelling/disaggregating approach. For example, the specific "harmonic content signature", reflecting the presence and strength of particular periodic-modes, might be used for the identification and modelling of different customer-classes in the total demand.
- Investigation of the combined diurnal and seasonal distributions and load compositions, with respect to the ratio of active/reactive power, can be used for various power quality (e.g. volt/var control) and voltage stability (e.g. voltage-dependency of demands) studies.
- Investigation of the possible connection between the rate of change of demands and daily and seasonal probability distributions of faults and short/long system interruptions can be used for devising reliability-based DSM, i.e. to perform suitable load/demand changes which will reduce occurrence of faults.
- Development of a methodology for generating electricity demand profiles, based on the identified consumption patterns from substations supplying varying percentage-contributions of various customer-classes. Initial results from Chapter 5 indicate very good agreement between modelled and estimated percentages. This approach can be developed to account for diurnal, as well as for seasonal demand variability, and can potentially include reactive power demand, thus generating dynamic profiles that can be used for a number of load modelling studies.
- The presented load profiling approaches can be expanded with more representative sets of measured demands in order to update some of the commonly used, but now outdated (e.g. due to changes in end-use electrical equipment) load profiles of different customer classes.
- Development of the load-disaggregation approach that can take into account voltage variations, therefore allowing for the expansion of the presented statistical analysis to

include static and dynamic load models, with correctly represented voltage-dependency of active/reactive power demands for different load types.

- Development of more comprehensive and more flexible load forecasting methodologies, that can take into account the results from the two load disaggregation chapters. From the DSM point of view, it would be particularly beneficial to investigate the potential for forecasting demands of specific load-types, rather than total aggregate demands and similarly for determining expected demands from different customer-classes at a sub-daily resolution (e.g. for different hours/half-hours of the day).

Bibliography

[1] C. W. Johnson, "What are emergent properties and how do they affect the engineering of complex systems," *Reliability in Engineering System Safety*, vol. 91, no. 12, pp. 1475-1481, 2006.

[2] S. D. Ramchurn, P. Vytelingum, A. Rogers and N. R. Jennings, "Putting the 'Smarts' into the Smart Grid: a Grand Challenge for Artificial Intelligence," *Communications of the ACM*, vol. 55, no. 4, 2012.

[3] National Grid, "National Grid Technical Specification - Substations," 1995.

[4] National Grid, "National Electricity Transmission System Performance Report," 2016.

[5] M. Dunk, "Engineering Standards - Grid & Primary Substation Design Guide," UK Power Networks, 2015.

[6] M. Dunk, "Engineering Design Standards - Secondary Substation Civil Design Standard," UK Power Networks, 2015.

[7] European Academies Science Advisory Council (EASAC) , "Transforming Europe's Electricity Sypply - An Infrastructure Strategy for Reliable, Renewable and Secure Power System," 2009.

[8] Ofgem, "Electricity Distribution Quality of Service Report," 2009.

[9] "Digest of United Kingdom Energy Statistics (DUKES)," Department for Business, Energy and Industrial Strategy, 2016.

[10] Elexon, "Transmission Losses," 2013.

[11] The World Bank, "Electric power transmission and distribution losses," [Online]. Available: http://data.worldbank.org/indicator/EG.ELC.LOSS.ZS. [Accessed 03 12 2016].

[12] Ofgem, "Electricity Security of Supply," 2015.

[13] H. Farhangi, "The path of the smart grid," *IEEE Power and Energy Magazine,* vol. 8, no. 1, pp. 18-28, 2010.

[14] J. P. Lopes, N. Hatziargyriou, J. Mutale, P. Djapic and N. Jenkins, "Integrating distributed generation into electric power systems: A review of drivers, challenges and opportunities," *Electric Power Systems Research,* vol. 77, no. 9, pp. 1189-1203, 2007.

[15] J. Driesen and R. Belmans, "Distributed Generation: Challenges and Possible Solutions," *2006 IEEE Power Engineering Society General Meeting*, 2006.

[16] J. M. Carrasco, L. G. Franquelo, J. T. Bialasiewicz, E. Galvan, R. Portillo, M. M. Prats, J. I. Leon and N. Moreno, "Power-Electronic Systems for the Grid Integration of Renewable Energy Sources: A Survey," *IEEE Transactions on Industrial Electronics,* vol. 53, no. 4, pp. 1002-1016, 2006.

[17] 1547-2003, IEEE, "Standard for Interconnecting Distributed Resources with Electric Power Systems," IEEE Std..

[18] Energy and Climate Change Commitee, "Low Carbon Network Infrastructure," House of Commons, 2016.

[19] US Department of Energy, "Grid Modernization Multi-Year Program Plan," US Department of Energy, 2015.

[20] H. Gharazi and R. Ghafurian, "Smart Grid: The Electricity Energy System of the Future," *Proceedings of the IEEE,* vol. 99, no. 6, pp. 917-921, 2011.

[21] "Smart Grid," US Department of Energy: Office of Electricity Delivery & Energy Reliability, [Online]. Available: https://energy.gov/oe/services/technology-development/smart-grid.

[22] I. Lampropoulos, W. L. Kling, P. F. Ribeiro and J. v. d. Berg, "History of demand side management and classification of demand response control schemes," in *IEEE Power & Energy Society General Meeting*, 2013.

[23] P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems and Smart Loads," *IEEE Transactions on Industrial Informatics,* vol. 7, no. 3, pp. 381-388, 2011.

[24] S. Djokic, C. Cresswell and A. Collin, "The Future of Residential Lighting: Shift from Incandescent to CFL to LED Light Sources," in *IES Annual Conference,* 2009.

[25] Department of Energy and Climate Change, "Energy Efficient Products - Helping us Cut Energy Use," Department of Energy and Climate Change, 2014.

[26] M. Binswanger, "Technological progress and sustainable development: what about the rebound effect?," *Ecological Economics,* vol. 36, no. 1, pp. 119-132, 2001.

[27] T. Barker, P. Ekins and T. Foxon, "The macro-economic rebound effect and the UK economy," *Energy Policy,* vol. 35, no. 10, pp. 4935-4946, 2007.

[28] E. Alfredsson, ""Green" Consumption - No Solution for Climate Change," *Energy,* vol. 29, no. 4, pp. 513-524, 2004.

[29] B. Kirby, "Spinning Reserve from Responsive Loads," Oak Ridge National Laboratory, 2003.

[30] L. Gelazanskas and K. A. Gamage, "Demand Side Management in Smart Grids: A Review and Proposals for Future Directions," *Sustainable Cities and Society,* vol. 11, pp. 22-30, 2014.

[31] C. Gellings, "The concept of demand-side management for electric utilities," *Proceedings of the IEEE,* vol. 73, no. 10, pp. 1468-1470, 2005.

[32] D. G. Infield, J. Short, C. Horne and L. L. Freris, "Potential for Domestic Dynamic Demand-Side Management in the UK," in *Power Engineering Society General Meeting,* 2007.

[33] National Grid, "Transmission Licence Standard Condition C17".

[34] "Energy Price Statistics," UK Government, [Online]. Available: https://www.gov.uk/government/collections/energy-price-statistics.

[35] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Optimal Residential Load Control With Price Prediction in Real-Time Electricity Pricing Environments," *IEEE Transactions on Smart Grid,* vol. 1, no. 2, pp. 120-133, 2010.

[36] P. Vytelingum, S. D. Ramchurn, T. D. Voice, A. Rogers and N. R. Jennings, "Trading Agents for the Smart Electricity Grid," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 2010.

[37] ELEXON, "The Electricity Trading Arrangements - A Beginner's Guide," ELEXON, 2015.

[38] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober and A. Leon-Garcia, "Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid," *IEEE Transactions on Smart Grid,* vol. 1, no. 3, 2010.

[39] T. Logenthiran, D. Srinivasan and T. Z. Shun, "Demand Side Management in Smart Grid Using Heuristic Optimization," *IEEE Transactions on Smart Grid,* vol. 3, no. 3, pp. 1244-1252, 2012.

[40] A. Molderink, V. Bakker, M. G. C. Bosman, J. L. Hurink and G. J. M. Smit, "Management and Control of Domestic Smart Grid Technology," *IEEE Transactions on Smart Grid,* vol. 1, no. 2, pp. 109-119, 2010.

[41] G. Strbac, "Demand Side Management: Benefits and Challenges," *Energy Policy,* vol. 36, no. 12, pp. 4419-4426, 2008.

[42] I. Atzeni, L. G. Ordóñez, G. Scutari, D. P. Palomar and J. R. Fonollosa, "Demand-Side Management via Distributed Energy Generation and Storage Optimization," *IEEE Transactions on Smart Grid,* vol. 4, no. 2, pp. 866-876, 2013.

[43] J. Aghaei and M.-I. Alizadeh, "Demand Response in Smart Electricity Grids Equipped with Renewable Energy Sources: A Review," *Renewable and Sustainable Energy Reviews,* vol. 18, pp. 64-72, 2013.

[44] C. Cecati, C. Citro and P. Siano, "Combined Operations of Renewable Energy Systems and Responsice Demand in a Smart Grid," *IEEE TRANSACTIONS ON SUSTAINABLE ENERGY,* vol. 2, no. 4, pp. 468-476, 2011.

[45] F. Rahimi and A. Ipakchi, "Demand Response as a Market Resource Under the Smart Grid Paradigm," *IEEE Transactions on Smart Grid,* vol. 1, no. 1, pp. 82-88, 2010.

[46] K. Clement-Nyns, E. Haesen and J. Driesen, "The Impact of Charging Plug-In Hybrid Electric Vehicles on a Residential Distribution Grid," *IEEE Transactions on Power Systems,* vol. 25, no. 1, pp. 371-380, 2010.

[47] S. Deilami, A. S. Masoum, P. S. Moses and M. A. S. Masoum, "Real-Time Coordination of Plug-In Electric Vehicle Charging in Smart Grids to Minimize Power Losses and Improve Voltage Profile," *IEEE Transactions on Smart Grid,* vol. 2, no. 3, pp. 456-467, 2011.

[48] Department of Energy & Climate Change, "Smart Meters - Quarterly Report," Department of Energy & Climate Change, 2016.

[49] A. Faruquia, D. Harrisb and R. Hledika, "Unlocking the €53 billion savings from smart meters in the EU: How increasing the adoption of dynamic tariffs could make or break the EU's smart grid investment," *Energy Policy,* vol. 38, no. 10, pp. 6222-6231, 2010.

[50] G. R. Newsham and B. G. Bowker, "The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: A review," *Energy Policy,* vol. 38, no. 7, pp. 3289-3296, 2010.

[51] Department for Business, Energy & Industrial Strategy, "Smart Meters: A Guide," UK Government, 2013. [Online]. Available: https://www.gov.uk/guidance/smart-meters-how-they-work.

[52] International Energy Agency, "World Energy Outlook - 2016," International Energy Agency, Paris, 2016.

[53] "Smart Meter Roll-Out Cost-Benefit Analysis," Department for Business, Energy & Industrial Strategy, 2016.

[54] National Statistics, "UK Energy Statistics - 2015," Department of Energy and Climate Change, 2015.

[55] J. E. Payne, "A survey of the electricity consumption-growth literature," *Applied Energy,* vol. 87, no. 3, pp. 723-731, 2010.

[56] A. Cavoukian, J. Polonetsky and C. Wolf, "SmartPrivacy for the Smart Grid: embedding privacy into the design of electricity conservation," *Identity in the Information Society,* vol. 3, no. 2, pp. 275-294, 2010.

[57] M. A. Lisovich, D. K. Mulligan and S. B. Wicker, "Inferring Personal Information from Demand-Response Systems," *IEEE Security & Privacy,* vol. 8, no. 1, 2010.

[58] E. L. Quinn, "Privacy and the New Energy Infrastructure," Social Science Research Network, Working Paper Series, 2009.

[59] O. Kosut, L. Jia, R. J. Thomas and L. Tong, "Malicious Data Attacks on Smart Grid State Estimation: Attack Strategies and Countermeasures," in *Smart Grid Communications (SmartGridComm),* 2010.

[60] X. Li, X. Liang, R. Lu, X. Shen, X. Lin and H. Zhu, "Securing smart grid: cyber attacks, countermeasures and challenges," *IEEE Communications Magazine,* vol. 50, no. 8, 2012.

[61] Z. Zhu, J. Tang, S. Lambotharan, W. H. Chin and Z. Fan, "An integer linear programming based optimization for home demand-side management in smart grid," in *Innovative Smart Grid Technologies (ISGT),* 2012.

[62] D. Niyato, L. Xiao and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine,* vol. 49, no. 4, 2011.

[63] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati and G. P. Hancke, "Smart Grid Technologies: Communication Technologies and Standards," *IEEE Transactions on Industrial Informatics,* vol. 7, no. 4, pp. 529-539, 2011.

[64] Z. Fan, P. Kulkarni, S. Gormus, C. Efthymiou, G. Kalogridis, M. Sooriyabandara, Z. Zhu, S. Lambotharan and W. H. Chin, "Smart Grid Communications: Overview of Research Challenges, Solutions, and Standardization Activities," *IEEE Communications Surveys & Tutorials,* vol. 15, no. 1, pp. 21-38, 2013.

[65] A. Monti and F. Ponci, "Power Grids of the Future: Why Smart Means Complex," in *Complexity in Engineering,* 2010.

[66] K. Moslehi and R. Kumar, "A Reliability Perspective of the Smart Grid," *IEEE Transactions on Smart Grid,* vol. 1, no. 1, pp. 57-64, 2010.

[67] A. Paisios, A. Ferguson and S. Z. Djokic, "Solar Analemma for Assessing Variations in Electricity Demands at MV Buses," in *Med Power 2016*, 2016.

[68] A. Paisios and S. Djokic, "Load Decomposition and Profiling for "Smart Grid" Demand-Side Management Applications," in *Data Analytics for Renewable Energy (DARE)*, Prague, 2013.

[69] A. Paisios and S. Djokic, "Decomposition of Aggregate Electricity Demand into the Seasonal-Thermal Components for Demand-Side Management Applications in "Smart Grids"," in *Lecture Notes in Computer Science : Chapter: Data Analytics for Renewable Energy Integration*, Springer, 2017, pp. 116-135.

[70] *MATLAB version: 9.0.0.341360 (R2016a),* Natick, Massachusetts: MathWorks Inc..

[71] A. J. Collin, "Advanced Load Modelling for Power System Studies - PhD Thesis," The University of Edinburgh, 2013.

[72] J. V. Milanovic, K. Yamashita, S. M. Villanueva, S. Ž. Djokic and L. M. Korunović, "International Industry Practice on Power System Load Modeling," *IEEE Transactions on Power Systems,* vol. 28, no. 3, pp. 3038-3046, 2013.

[73] C. Cresswell, S. Djokic, K. Ochije and E. Macpherson, "Modelling of non-linear electronic loads for power system studies: a qualitative approach," in *19th International Conference on Electricity Distribution*, Vienna, 2007.

[74] C. Cresswell and S. Djokic, "Representation of Directly Connected and Drive-controlled Induction Motors.," in *International Conference on Electrical Machines*, 2008.

[75] C. Cresswell and S. Djokic, "Steady-state models of low energy consumption light sources," in *16th Power Systems Computation Conference*, 2008.

[76] A. J. Collin, G. Tsagarakis, A. E. Kiprakis and S. McLaughlin, "Development of Low-Voltage Load Models for the Residential Load Sector," *IEEE Transactions on Power Systems,* vol. 29, no. 5, pp. 2180-2188, 2014.

[77] I. Richardsona, M. Thomsona and D. Infieldb, "A high-resolution domestic building occupancy model for energy demand simulations," *Energy and Buildings,* vol. 40, no. 8, pp. 1560-1566, 2008.

[78] A. Capasso, W. Grattieri, R. Lamedica and A. Prudenzi, "A bottom-up approach to residential load modeling," *IEEE Transactions on Power Systems,* vol. 9, no. 2, pp. 957 - 964, 1994.

[79] J. V. Paatero and P. D. Lund, "A model for generating household electricity load profiles," *Energy Reseasrch,* vol. 30, no. 5, pp. 273-290, 2006.

[80] M. Kavgica, A. Mavrogiannia, D. Mumovica, A. Summerfieldb, Z. Stevanovicc and M. Djurovic-Petrovicd, "A review of bottom-up building stock models for energy consumption in the residential sector," *Building and Environment,* vol. 45, no. 7, pp. 1683-1697, 2010.

[81] A. Collin, I. Hernando-Gil and J. Acosta, "An 11 kV steady state residential aggregate load model. Part 1: Aggregation methodology," in *PowerTech - IEEE*, Trondheim, 2011.

[82] K. Yamashita, S. Djokic, J. Matevosyan, F. Resende, L. Korunovic, Z. Dong and J. Milanovic, "Modelling and Aggregation of Loads in Flexible Power Networks," *IFAC Proceedings Volumes,* vol. 45, no. 12, p. 405–410, 2012.

[83] D. P. Stojanović, L. M. Korunović and J. Milanović, "Dynamic load modelling based on measurements in medium voltage distribution network," *Electric Power Systems Research,* vol. 78, no. 2, pp. 228-238, 2008.

[84] B.-K. Choi, H.-D. Chiang, Y. Li, H. Li, Y.-T. Chen, D.-H. Huang and M. Lauby, "Measurement-based dynamic load models: derivation, comparison, and validation," *IEEE Transactions on Power Systems,* vol. 21, no. 3, pp. 1276-1283, 2006.

[85] J. Ma, R.-m. He, Z.-y. Dong and D. J. Hill, "Measurement-based Load Modeling using Genetic Algorithms," in *IEEE Congress on Evolutionary Computation*, 2007.

[86] A. A. d. Silva, C. Ferreira and G. Torres, "Dynamic load modeling based on a nonparametric ANN," in *Intelligent Systems Applications to Power Systems*, 1996.

[87] A. Stankovic and B. Lesieutre, "Parametric variations in dynamic models of induction machine clusters," *IEEE Transactions on Power Systems,* vol. 12, no. 4, pp. 1549-1554, 2002.

[88] "Elexon UK," Elexon , [Online]. Available: https://www.elexon.co.uk/.

[89] Elexon, "Load Profiles and their Use in Electricity Settlement," Elexon, 2013.

[90] R. Li, C. Gu, Y. Zhang and F. Li, "Implementation of load profile test for electricity distribution networks," in *IEEE Power and Energy Society General Meeting*, 2012.

[91] E. González-Romera, M. Jaramillo-Morán and D. Carmona-Fernández, "Monthly electric energy demand forecasting with neural networks and Fourier series," *Energy Conversion and Management,* vol. 49, no. 11, pp. 3135-3142, 2008.

[92] S. Moutter, P. Bodger and P. Gough, "Spectral decomposition and extrapolation of variations in electricity loading," *IEE Proceedings C - Generation, Transmission and Distribution,* vol. 133, no. 5, pp. 247 - 255, 1986.

[93] F. McLoughlin, A. Duffy and M. Conlon, "Evaluation of time series techniques to characterise domestic electricity demand," *Energy,* vol. 50, pp. 120-130, 2013.

[94] L. Magnano and J. Boland, "Generation of synthetic sequences of electricity demand: Application in South Australia," *Energy,* vol. 32, no. 11, p. 2230–2243, 2007.

[95] Y. Ji, P. Xu and Y. Ye, "HVAC terminal hourly end-use disaggregation in commercial buildings with Fourier series model," *Energy and Buildings,* vol. 97, pp. 33-46, 2015.

[96] A. J. Collin, G. Tsagarakis, A. E. Kiprakis and S. McLaughlin, "Multi-scale electrical load modelling for demand-side management," in *3rd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe)*, 2012.

[97] J. Palmer, N. Terry and T. Kane, "Further Analysis of the Household Electricity Survey - Early Findings: Demand Side Management," Department of Energy & Climate Change, 2013.

[98] M. Dolman, I. Walker, A. Wright and G. Stuart, "Demand Side Response in the Non-Domestic Sector," Element Energy Limited, 2012.

[99] G. Franco and A. H. Sanstad, "Climate change and electricity demand in California," *Climate Change,* vol. 87, pp. 139-151, 2008.

[100] A. D. Amato, M. Ruth, P. Kirshen and J. Horwitz, "Regional Energy Demand Responses To Climate Change: Methodology And Application To The Commonwealth Of Massachusetts," *Climate Change,* vol. 71, no. 1, pp. 175-201, 2005.

[101] K. Pilli-Sihvolaa, P. Aatolab, M. Ollikainenb and H. Tuomenvirtaa, "Climate change and electricity consumption—Witnessing increasing or decreasing use and costs?," *Energy Policy,* vol. 38, no. 5, pp. 2409-2419, 2010.

[102] J. Squalli, "Electricity consumption and economic growth: Bounds and causality analyses of OPEC members," *Energy Economics,* vol. 29, no. 6, pp. 1192-1205, 2007.

[103] Y. Wolde-Rufael, "Electricity consumption and economic growth: a time series experience for 17 African countries," *Energy Policy,* vol. 34, no. 10, pp. 1106-1114, 2006.

[104] L. Suganthi and A. A. Samuel, "Energy models for demand forecasting - A review," *Renewable and Sustainable Energy Reviews,* vol. 16, no. 2, pp. 1223-1240, 2012.

[105] J. M. Griffin, "The Effects of Higher Prices on Electricity Consumption," *The Bell Journal of Economics and Management Science,* vol. 5, no. 2, pp. 515-539, 1974.

[106] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts and C. Evans, "Household Electricity Survey - A Study of Domestic Electricity Product Usage," Intertek, 2012.

[107] B. Psiloglou, C. Giannakopoulos, S. Majithia and M. Petrakis, "Factors affecting electricity demand in Athens, Greece and London, UK: A comparative assessment," *Energy,* vol. 34, no. 11, pp. 1855-1863, 2009.

[108] C.-L. Hor, S. Watson and S. Majithia, "Analyzing the impact of weather variables on monthly electricity demand," *IEEE Transactions on Power Systems,* vol. 20, no. 4, pp. 2078-2085, 2005.

[109] A. Albert, T. Gebru, J. Ku, J. Kwac, J. Leskovec and R. Rajagopal, "Drivers of Variability in Energy Consumption," in *ECML-PKDD DARE Workshop on Energy Analytics*, 2013.

[110] M. Santamouris, N. Papanikolaou, I. Livada, I. Koronakis, C. Georgakis, A. Argiriou and D. Assimakopoulos, "On the impact of urban climate on the energy consumption of buildings," *Solar Energy*, vol. 70, no. 3, pp. 201-216, 2001.

[111] B. P. Hayes, "PhD Thesis: Distributed Generation and Demand Side Managment: Applications to Transmission System Operation," The Unicersity of Edinburgh, 2013.

[112] G. Sinden, "Characteristics of the UK wind resource: Long-term patterns and relationship to electricity demand," *Energy Policy*, vol. 35, no. 1, pp. 112-127, 2007.

[113] J. Moral-Carcedo and J. Vicéns-Otero, "Modelling the non-linear response of Spanish electricity demand to temperature variations," *Energy Economics*, vol. 27, no. 3, pp. 477-494, 2005.

[114] M. Davies, "The relationship between weather and electricity demand," *Proceedings of the IEE - Part C: Monographs*, vol. 106, no. 9, pp. 27-37, 1959.

[115] M. Bessec and J. Fouquau, "The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach," *Energy Economics*, vol. 30, no. 5, pp. 2705-2721, 2008.

[116] G. K. Tso and K. K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761-1768, 2007.

[117] M. Ranjan and V. Jain, "Modelling of electrical energy consumption in Delhi," *Energy*, vol. 24, pp. 351-361, 1999.

[118] Z. Ismail, F. Jamaluddin and F. Jamaludin, "Time Series Regression Model for Forecasting Malaysian Electricity Load Demand," *Asian Journal of Mathematics & Statistics*, vol. 1, no. 3, pp. 139-149, 2008.

[119] H. Mousazadeha, A. Keyhania, A. Javadi, H. Mobli, K. Abrinia and A. Sharifi, "A review of principle and sun-tracking methods for maximizing solar systems output," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 1800-1818, 2009.

[120] K. Chong and C. Wong, "General formula for on-axis sun-tracking system and its application in improving tracking accuracy of solar collector," *Solar Energy*, vol. 83, no. 3, pp. 298-305, 2009.

[121] D. H. Vu, K. M. Muttaqi and A. P. Agalgaonkar, "Short-term load forecasting using regression based moving windows with adjustable window-sizes," in *IEEE Industry Applications Society Annual Meeting*, 2014.

[122] H. Al-Hamadi and S. Soliman, "Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model," *Electric Power Systems Research*, vol. 68, no. 1, pp. 47-59, 2004.

[123] V. Dordonnat, S. J. Koopman and M. Ooms, "Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling," *Computational Statistics & Data Analysis*, vol. 56, no. 11, pp. 3134-3152, 2012.

[124] A. Mutanen, M. Ruska, S. Repo and P. Jarventausta, "Customer Classification and Load Profiling Method for Distribution Systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755-1763, 2011.

[125] D. Gerbec, S. Gasperic, I. Smon and F. Gubina, "An approach to customers daily load profile determination," in *IEEE Power Engineering Society Summer Meeting*, 2002.

[126] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1232-1239, 2004.

[127] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas and M. Kolehmainena, "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," *Applied Energy*, vol. 87, no. 11, pp. 3538-3545, 2010.

[128] "Sub-national electricity and gas consumption statistics," Department of Energy and Climate Change, 2015.

[129] G. Nourbakhsh, G. Eden, D. McVeigh and A. Ghosh, "Chronological Categorization and Decomposition of Customer Loads," *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 2270-2277, 2012.

[130] S. Ramos and Z. Vale, "Data Mining techniques to support the classification of MV electricity customers," in *IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, 2008.

[131] C. Walton and S. Carter, "Secondary substation load profiling — Identification and visualisation of changes," in *Conference and Exhibition on Electricity Distribution (CIRED 2013), 22nd International*, 2013.

[132] J. Jardini, C. Tahan, M. Gouvea, S. Ahn and F. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," in *IEEE Transactions on Power Delivery*, 2002.

[133] G. Chicco, R. Napoli and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933-940, 2006.

[134] I. Prahastono, D. King and C. S. Ozveren, "A review of Electricity Load Profile Classification methods," in *42nd International Universities Power Engineering Conference*, 2007.

[135] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241-254, 1967.

[136] W. Shen, W. Wu, H. Sun, B. Zhang and M. Jiang, "A wave filtering based electric load curve decomposition method for AGC," in *International Conference on Power System Technology (POWERCON)*, 2010.

[137] S. D. Ramchurn, P. Vytelingum, A. Rogers and N. Jennings, "Agent-based control for decentralised demand side management in the smart grid," in *10th International Conference on Autonomous Agents and Multiagent Systems*, 2011.

[138] J. T. Wenders, "Peak Load Pricing in the Electric Utility Industry," *The Bell Journal of Economics*, vol. 7, no. 1, pp. 232-241, 1976.

[139] A. Salimi-beni, D. Farrokhzad, M. Fotuhi-Firuzabad and S. J. Alemohammad, "A New Approach to Determine Base, Intermediate and Peak-Demand in an Electric Power System," in *International Conference on Power System Technology*, 2006.

[140] E. Mayhorn, R. Butner, M. Baechler, G. Sullivan and H. Hao, "Characteristics and Performance of Existing Load Disaggregation Technologies," Pacific Northwest National Laboratory, 2015.

[141] G. Hart, "Residential energy monitoring and computerized surveillance via utility power flows," *IEEE Technology and Society Magazine* , vol. 8, no. 2, pp. 12-16, 2002.

[142] M. Zeifman and K. Roth, "Nonintrusive Appliance Load Monitoring: Review and Outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76-84, 2011.

[143] K. C. Armela, A. Guptab, G. Shrimalic and A. Albert, "Is disaggregation the holy grail of energy efficiency? The case of electricity," *Energy Policy*, vol. 52, pp. 213-234, 2013.

[144] S. Ng, J. Liang and J. Cheng, "Automatic appliance load signature identification by statistical clustering," in *8th International Conference on Advances in Power System Control, Operation and Management (APSCOM)*, 2009.

[145] L. Farinaccio and R. Zmeureanu, "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses," *Energy and Buildings*, vol. 30, no. 3, pp. 245-259, 1999.

[146] T. Onoda, H. Murata, G. Ratsch and K. Muller, "Experimental analysis of support vector machines with different kernels based on non-intrusive monitoring data," in *Proceedings of the 2002 International Joint Conference on Neural Networks*, 2002.

[147] A. G. Ruzzelli, C. Nicolas, A. Schoofs and G. M. P. O'Hare, "Real-Time Recognition and Profiling of Appliances through a Single Electricity Sensor," in *7th Annual IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks (SECON)*, 2010.

[148] M. Zeifman, "Disaggregation of home energy display data using probabilistic approach," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 1, 2012.

[149] O. Parson, "Disaggregated Homes," 2012. [Online]. Available: http://blog.oliverparson.co.uk/2012/06/public-data-sets-for-nialm.html.

[150] BigSofa, "Understanding the customer experience of dynamically teleswitched (DTS) meters and tariffs," Ofgem, 2014.

[151] Ofgem, "Standard Conditions of the Electricity Distribution Licence," Ofgem, 2017.

[152] Ofgem, "Open letter to DNOs and other interested parties - household electricity smart-meter data - privacy plans," [Online]. Available: https://www.ofgem.gov.uk.

[153] J. D. Hobby, A. Shoshitaishvili and G. H. Tucci, "Analysis and Methodology to Segregate Residential Electricity Consumption in Different Taxonomies," *IEEE Transactions on Smart Grid,* vol. 3, no. 1, pp. 217-224, 2012.

[154] S. Lee and J.-W. Park, "A Reduced Multivariate Polynomial Model for Estimation of Electric Load Composition," in *IEEE Industry Applications Conference*, 2007.

[155] J. Duan, D. Czarkowski and Z. Zabar, "Neural network approach for estimation of load composition," in *Proceedings of the 2004 International Symposium on Circuits and Systems*, 2004.

[156] Y. Xu and J. V. Milanovic, "Estimation of percentage of controllable load in total demand at bulk supply points," in *Proc. 9th Med. Conference on Power Generation, Transmission, Distribution and Energy Conversion,* Athens, 2014.

[157] Y. Xu and J. V. Milanović, "Artificial-Intelligence-Based Methodology for Load Disaggregation at Bulk Supply Point," *IEEE Transactions on Power Systems,* vol. 30, no. 2, pp. 795-803, 2015.

[158] Y. Xu and J. V. Milanović, "Day-Ahead Prediction and Shaping of Dynamic Response of Demand at Bulk Supply Points," *IEEE Transactions on Power Systems,* vol. 31, no. 4, pp. 3100-3108, 2016.

[159] M. Isaac and D. P. v. Vuuren, "Modeling global residential sector energy demand for heating and air conditioning in the context of climate change," *Energy Policy*, vol. 37, no. 2, pp. 507-521, 2009.

[160] Y. Goude, R. Nedellec and N. Kong, "Local Short and Middle Term Electricity Load Forecasting with Semi-Parametric Additive Models," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 440-446, 2014.

[161] T. Yalcinoz and U. Eminoglu, "Short term and medium term power distribution load forecasting by neural networks," *Energy Conversion and Management,* vol. 46, no. 9-10, pp. 1394-1405, 2005.

[162] N. Abu-Shikhah, F. Elkarmi and O. M. Aloquili, "Medium-Term Electric Load Forecasting Using Multivariable Linear and Non-Linear Regression," *Smart Grid and Renewable Energy,* vol. 2, pp. 126-135, 2011.

[163] H. Al-Hamadi and S. Soliman, "Long-term/mid-term electric load forecasting based on short-term correlation and annual growth," *Electric Power Systems Research,* vol. 74, no. 3, pp. 353-361, 2005.

[164] J. W. Taylor, L. M. d. Menezes and P. E. McSharry, "A comparison of univariate methods for forecasting electricity demand up to a day ahead," *International Journal of Forecasting,* vol. 22, no. 1, pp. 1-16, 2006.

[165] K. Metaxiotis, A. Kagiannas, D. Askounis and J. Psarras, "Artificial intelligence in short term electric load forecasting: A state-of-the-art survey for the researcher," *Energy Conversion and Management,* vol. 44, no. 9, pp. 1525-1534, 2003.

[166] H. Hahn, S. Meyer-Nieberg and S. Pickl, "Electric load forecasting methods: Tools for decision making," *European Journal of Operational Research,* vol. 199, no. 3, pp. 902-907, 2009.

[167] H. Alfares and N. Mohammad, "Electric load forecasting: Literature survey and classification of methods," *International Journal of Systems Science,* vol. 33, no. 1, pp. 23-34, 2002.

[168] L. Korunovic, D. Stojanovic and J. Milanovic, "Identification of static load characteristics based on measurements in medium-voltage distribution network," *IET Generation, Transmission & Distribution,* vol. 2, no. 2, pp. 227-234, 2008.

[169] A. Azadeh, M. Saberi, S. Ghaderi, A. Gitiforouz and V. Ebrahimipour, "Improved estimation of electricity demand function by integration of fuzzy system and data mining approach," *Energy Conversion and Management,* vol. 49, pp. 2165-2177, 2008.

[170] G. J. Tsekouras, N. D. Hatziargyriou and E. N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," *IEEE Transactions on Power Systems,* vol. 22, no. 3, pp. 1120-1128, 2007.

[171] K. N. Hasan, J. V. Milanović, P. Turner and V. Turnham, "A Step-by-Step Data Processing Guideline for Load Model Development Based on Field Measurements," in *IEEE PowerTech,* Eindhoven, 2015.

[172] P.-F. Pai and W.-C. Hong, "Support vector machines with simulated annealing algorithms in electricity load forecasting," *Energy Conversion and Management,* vol. 46, no. 17, pp. 2669-2688, 2005.

[173] J.-F. Chen, W.-M. Wang and C.-M. Huang, "Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting," *Electric Power Systems Research,* vol. 34, no. 3, pp. 187-196, 1995.

[174] S. Yao, Y. Song, L. Zhang and X. Cheng, "Wavelet transform and neural networks for short-term electrical load forecasting," *Energy Conversion and Management,* vol. 41, no. 18, pp. 1975-1988, 2000.

[175] C. V. Loan, Computational frameworks for the fast Fourier transform, Society for Industrial and Applied Mathematics Philadelphia, 1992.

[176] "Voltage Limits Assessment Discussion Paper," Western Power Distribution - Network Equilibrium, 2016.

[177] C. Chatfield, The Analysis of Time Series: An Introduction, London: Chapman and Hall, 1989.

[178] D. W. Ricker, "Stochastic Processes," in *Echo Signal Processing,* Springer, 2003, pp. 23-28.

[179] C. A.D. and J. Ord, *Spatial Autocorrelation,* London: Pion, 1973.

[180] R. Hubbard, " Blurring the Distinctions Between p's and a's in Psychological Research," *Theory Psychology,* vol. 12, no. 3, pp. 295-327, 2004.

[181] E.W. Weisstein, "P-Value," A Wolfram Web Resource, 2016. [Online]. Available: http://mathworld.wolfram.com/P-Value.html.

[182] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician,* vol. 70, no. 2, pp. 129-133, 2016.

[183] S. Goodman, "A Dirty Dozen: Twelve P-Value Misconceptions," *Interpretation of Quantitative Research,* vol. 45, no. 3, pp. 135-140, 2008.

[184] G. W. Snedecor and W. G. Cochran, in *Statistical Methods,* 1967, p. 321.

[185] M. R. Harwell, E. N. Rubinstein, W. S. Hayes and C. C. Olds, "Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases," *Journal of Educational and Behavioral Statistics,* vol. 17, no. 4, 1992.

[186] G. V. Glass, P. D. Peckham and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Review of Educational Research,* vol. 42, no. 3, pp. 237-288, 1972.

[187] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association,* vol. 47, no. 260, 1952.

[188] G. Tsagarakis, A. J. Collin and A. E. Kiprakis, "Modelling the electrical loads of UK residential energy users," in *47th International Universities Power Engineering Conference (UPEC)*, 2012.

[189] I.-S. Ilie, I. Hernando-Gil and S. Z. Djokic, "Theoretical interruption model for reliability assessment of power supply systems," *IET Generation, Transmission & Distribution,* vol. 8, no. 4, pp. 670-681, 2014.

[190] I. Hernando-Gil, *PhD Thesis: Integrated Assessment of Quality of Supply in Future Electricity Networks,* Edinburgh: The University of Edinburgh, 2014.

[191] "The National Meteorological Library and Archive - Fact Sheet: Observations," Met Office.

[192] M. Allaby, A Dictionary of Earth Sciences, Oxford: Oxford University Press, 2008.

[193] E. Raisz, "The Analemma," *Journal of Geography,* vol. 40, no. 3, 1941.

[194] I. Reda and A. Andreas, "Solar Position Algorithm for Solar Radiation Applications," *Solar Energy*, vol. 76, pp. 577-589, 2004.

[195] "MIDC SOLPOS Calculator," NREL - National Renewable Energy Laboratory, 2015. [Online]. Available: http://www.nrel.gov/midc/solpos/solpos.html.

[196] E. W. Weisstein, "Correlation Coefficient," MathWorld-A Wolfram Web Resource, [Online]. Available: http://mathworld.wolfram.com/CorrelationCoefficient.html.

[197] J. L. Mayers, A. D. Well and R. F. Lorch, Research Design and Statistical Analysis, New York: Routledge, 2010.

[198] J. D. Hamilton, "Linear Regression Models," in *Time Series Analysis*, Princeton University Press, pp. 200-231.

[199] E. Ostertagová, "Modelling using Polynomial Regression," *Procedia Engineering*, vol. 48, pp. 500-506, 2012.

[200] J. P. Barrett, "The Coefficient of Determination - Some Limitations," *The American Statistician*, vol. 28, no. 1, 1974.

[201] H. H. Harman, Modern Factor Analysis, University of Chicago Press, 1976.

[202] H. Schneeweiss and H. Mathes, "Factor Analysis and Principal Components," *Journal of Multivariate Analysis*, vol. 55, no. 1, pp. 105-124, 1995.

[203] D. D. Suhr, "Principal Component Analysis versus Exploratory Factor Analysis," in *SUGI 30*, 2005.

[204] K. Baba, R. Shibata and M. Sibuya, "Partial Correlation and Conditional Correlation as Measures of Conditional Independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657-664, 2004.

[205] A. Vargha, L. R. Bergman and H. D. Delaney, "Interpretation problems of the partial correlation with nonnormally distributed variables," *Quality & Quantity*, vol. 47, no. 6, pp. 3391-3402, 2013.

[206] D. Cramer, "A Cautionary Tale of Two Statistics: Partial Correlation and Standardized Partial Regression," *The Journal of Psychology Interdisciplinary and Applied*, vol. 137, no. 5, pp. 507-511, 2003.

[207] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, vol. 78, no. 368, pp. 829-836, 1979.

[208] W. S. Cleveland and S. J. Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596-610, 1988.

[209] S. Manabe and R. T. Wetherald, "Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity," *Journal of Atmospheric Sciences*, vol. 24, no. 3, 1967.

[210] X. Du, B. Li, H. Liu, D. Yang, W. Yu, J. Liao, Z. Huang and K. Xia, *The Response of Human Thermal Sensation and Its Prediction to Temperature Step-Change*, PLoS ONE 9(8), 2014.

[211] R. G. Steadman, "The Assessment of Sultriness. Part I: A Temperature-Humidity Index Based on Human Physiology and Clothing Science," *Journal of Applied Meteorology*, vol. 18, no. 7, pp. 861-873, 1979.

[212] Department for Business, Energy and Industrial Strategy, "Sub-National Electricity Consumption Data," UK Government, [Online]. Available: https://data.gov.uk/dataset/electricity_consumption_at_local_authority_level.

[213] "Distribution Long Term Development Statement," SP Distribution, 2011.

[214] T. E. Dielman, "Least Absolute Value Regression: Recent Contributions," *Journal of Statistical Computation and Simulation*, vol. 75, no. 4, pp. 263-286, 2005.

[215] O. Buyukalaca, H. Bulut and T. Yılmaz, "Analysis of variable-base heating and cooling degree-days for Turkey," *Applied Energy*, vol. 69, pp. 269-283, 2001.

[216] K. Papakostas and N. Kyriakis, "Heating and cooling degree-hours for Athens and Thessaloniki, Greece," *Renewable Energy*, vol. 30, pp. 1873-1880, 2005.

[217] E. Valor, V. Meneu and V. Caselles, "Daily Air Temperature and Electricity Load in Spain," *American Meteorological Society*, vol. 40, pp. 1413-1421, 2001.

[218] "Climate Monitoring - UKCP09," Met Office - UK Government, [Online]. Available: http://www.metoffice.gov.uk/climatechange/science/monitoring/ukcp09/faq.html.

[219] H. Abdi, "Partial Regression Coefficients," *Encyclopedia of Social Sciences Research Methods*, pp. 978-982, 2004.

[220] C. Cresswell, "PhD Thesis: Steady State Load Models for Power System Analysis," The University of Edinburgh, 2009.

[221] J. Liang, S. K. K. Ng, G. Kendall and J. W. M. Cheng, "Load Signature Study—Part I: Basic Concept, Structure and Methodology," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 551-560, 2010.

[222] A. F. A. El-Gawad, "Studying the impact of different lighting loads on both harmonics and power factor," in *42nd International Universities Power Engineering Conference*, 2007.